Zeman, F. (2025). Uso de corpus propios en la enseñanza del léxico formulaico en el aula de Español como Lengua Extranjera. *RILEX. Revista sobre investigaciones léxicas,* 8/I. pp. 123-153

USO DE CORPUS PROPIOS EN LA ENSEÑANZA DEL LÉXICO FORMULAICO EN EL AULA DE ESPAÑOL COMO LENGUA EXTRANJERA (ELE)

USE OF SELF-COMPILED CORPORA FOR FORMULAIC VOCABULARY TEACHING IN THE SPANISH AS A FOREIGN LANGUAGE CLASSROOM (SFL)

Filip Zeman
Universidad Autónoma de Madrid
filip.zeman@uam.es

Resumen

Este artículo propone e ilustra con ejemplos una nueva técnica de enseñanza-aprendizaje del español como lengua extranjera basada en el enfoque léxico enriquecido por la creación y consulta de corpus propios. La propuesta se centra a la enseñanza del léxico, en concreto del léxico pluriverbal: partiendo de los avances en la enseñanza de léxico de segundas lenguas y de las aportaciones de la lingüística de corpus, presenta una serie de actividades con los corpus propios diseñadas con diferentes herramientas del sistema de consulta y gestión de corpus *Sketch Engine* para demostrar el potencial didáctico de las estrategias propuestas.

Palabras clave: corpus propio, ELE, enfoque léxico, lenguaje formulaico, *Sketch Engine*, estudiantes avanzados.

Abstract

This article proposes and illustrates with examples a new Spanish as a foreign language teaching technique which is based on the Lexical Approach empowered by the creation and didactic use of self-compiled corpora. The proposed technique is aimed at the teaching of multiword expressions. Our proposition derives from the latest acknowledgements in the areas of second-language vocabulary teaching and corpus linguistics. We will present a series of activities which are created around the use of self-compiled corpora. The activities make use of the different tools of the corpus manager and text analysis software *Sketch Engine*. Subsequently, the article examines the didactic potential of such activities.

Keywords: self-compiled corpus, SFL, lexical approach, formulaic language, *Sketch Engine*, advanced learners.

Recibido: 01-05-2024
Aceptado: 20-11-2024

DOI: https://doi.org/10.17561/rilex.8.1.8906

1. INTRODUCCIÓN

En el ámbito del español como lengua extranjera, la enseñanza del léxico ha experimentado un desarrollo notable mediante la introducción de nuevas maneras de entender y enseñar el léxico, la diversificación de las tareas, la introducción del factor afectivo, el desarrollo del aprendizaje autónomo y la atención a la pluriverbalidad. Una de las propuestas recientes en este ámbito consiste en aplicar los corpus lingüísticos en el aula. Propuestas de su uso en la enseñanza aparecieron junto con el surgimiento de la lingüística computacional como disciplina, pero las funcionalidades limitadas de las herramientas de consulta no permitieron una adaptación eficiente hasta hace relativamente poco. Gracias al desarrollo tecnológico, los corpus modernos se han convertido en sistemas de fácil acceso y consulta, que permiten anotar los datos y realizar computaciones estadísticas complejas de forma automática, y que presentan los resultados de búsqueda de forma clara y resumida al usuario. El presente trabajo sigue esta línea para proponer una nueva técnica de enseñanza del léxico del español con el uso de corpus propios. Teniendo en cuenta los resultados de los recientes estudios sobre la adquisición del léxico y los beneficios de la diversificación metodológica, presentaremos esta nueva técnica junto con una muestra de actividades de clase basadas en el *enfoque léxico* de Michael Lewis. Con ello pretendemos contribuir a que se superen las reticencias sobre el uso directo de los corpus en el aula de lenguas extranjeras. Este trabajo expone las bases teóricas de un futuro estudio empírico, en el que la técnica de enseñanza presentada se pondrá en práctica y en el que se medirá su efectividad.

Para contextualizar el trabajo, resumiremos en la sección 2 algunos de los avances en la enseñanza de léxico de las últimas décadas e introduciremos la teoría del enfoque léxico. También presentaremos la noción de *corpus lingüístico*, y revisaremos la historia y las formas de su aplicación en la enseñanza de lenguas extranjeras. En la sección 3, describiremos las herramientas y funcionalidades del programa *Sketch Engine* usadas en esta propuesta y

presentaremos una muestra de actividades de aula basadas en la consulta de corpus propios. Finalmente, en la sección 4 resumiremos las conclusiones de este estudio.

2. CONSIDERACIONES TEÓRICAS Y REVISIÓN BIBLIOGRÁFICA

Para contextualizar nuestra propuesta, en esta sección presentaremos brevemente su marco teórico y revisaremos los trabajos más relevantes de cada campo temático.

2.1. El enfoque léxico de Michael Lewis

La metodología de la enseñanza del léxico se encuentra hoy en día en una fase consolidada gracias a los métodos comunicativos de enseñanza y, especialmente, gracias al *enfoque léxico*. Este enfoque, que introdujo el lingüista Michael Lewis hace 30 años, supuso un cambio esencial en la percepción del léxico dentro y fuera del aula de lenguas extranjeras. Por primera vez se tomó conciencia de la naturaleza interconectada de las palabras dentro de un idioma. Se prestó atención a las expresiones complejas o *pluriverbales*, incluidas dentro del concepto de *unidades léxicas* (o *chunks* en inglés), tal y como lo interpretó Lewis (2008, p. 7). Con el término *unidad léxica* denominamos las "secuencias de palabras, continuas o discontinuas, que parecen prefabricadas, es decir, que se almacenan y se recuperan de la memoria como un todo, en lugar de generarse desde cero cada vez que se producen o están sujetas a las reglas de la gramática" (Wray, 2002, p. 9). La validez de la teoría de Lewis se confirmó con la llegada de la lingüística computacional. Los corpus permitieron, por primera vez, realizar un análisis estadístico rápido de la expresión lingüística humana y, aunque las estimaciones de las diferentes investigaciones relevantes varían en la extensión, indican que hasta la mitad de nuestra producción diaria es formulaica (Wray, 2002), es decir, "está formada por cadenas de palabras que, por su recurrencia, se recuperan de la memoria como un bloque" (Rufat & Jiménez Calderón, 2017, p. 48). El concepto de lo formulaico es muy abarcador: no solo incluye

las categorías conocidas de unidades léxicas pluriverbales (como las frases hechas o las colocaciones), sino también cualquier combinación de palabras relativamente frecuente y estable.

La pluriverbalidad para un aprendiz de lenguas extranjeras (LE) es clave para su fluidez expresiva y un procesamiento rápido y eficiente del lenguaje (Lewis, 2008, p. 15) tal y como lo procesan los hablantes nativos. Al no estar ocupada en producir combinaciones óptimas de palabras, la capacidad reflexiva del cerebro humano se puede destinar a otras tareas cognitivas –construir el contenido del enunciado, planificar el discurso, etc.–. El almacenamiento de lexis en bloques, más que en palabras sueltas, también corresponde mejor a la estructura del *lexicón mental* (Lewis, 2008, p. 78), el sistema complejo que usamos para almacenar las palabras y para relacionarlas en el plano horizontal (mediante conexiones sinonímicas o antonímicas, por ejemplo) y vertical (mediante vínculos hiperonímicos, meronímicos, etc.). Por último, el conocimiento de los bloques léxicos facilita la aplicación correcta de los principios *de idiomaticidad* y *selección libre*.

Estos dos principios, introducidos por un contemporáneo de Lewis, John Sinclair, son los que rigen la selección léxica (1991, pp. 109-10). Según la tesis de la selección libre, el hablante dispone de libertad casi absoluta a la hora de hablar y elegir palabras y estructuras (y sus combinaciones); las únicas restricciones que existen son las gramaticales. Sin embargo, los factores idiomáticos restringen estas posibilidades combinatorias en el uso real de la lengua. Por ejemplo, para despedirnos solemos recurrir a expresiones comúnmente usadas en nuestra comunidad lingüística, aunque en un principio la gramática podría generar muchas otras. Los hablantes nativos cultos tienen interiorizadas estas reglas, pero los aprendientes de segundas lenguas no, incluso cuando tienen un conocimiento léxico avanzado, por lo que a menudo *suenan a extranjero*. Al centrar la atención de los aprendientes en la existencia de los bloques léxicos, el enfoque léxico permite superar esta limitación.

Los corpus lingüísticos, que son el medio de nuestra propuesta y que introduciremos en la siguiente sección, aparecen escasamente en el enfoque léxico. A primera vista, este hecho puede parecer contradictorio: el enfoque léxico enfatiza el uso de textos auténticos y los corpus son la fuente más extensa de textos auténticos. No obstante, como detallaremos en la siguiente sección, en la época de la creación del enfoque léxico los corpus estaban empezando a transformarse en digitales, por lo que sus herramientas y funciones aún eran limitadas. Cuando Lewis menciona los corpus, los describe como una fuente útil para crear obras didácticas y lexicográficas de calidad. En otras palabras, Lewis les concibió como una fuente útil para expertos encargados de crear materiales de referencia –manuales de lengua, diccionarios, libros de referencia– (2008, p. 206). En ningún momento plantea su uso como herramientas de consulta manejadas por los aprendientes. Como veremos, hoy en día los corpus han evolucionado en cuanto a la cantidad de su contenido y la facilidad con la se acceden, por lo que se han convertido en herramientas útiles para el aprendiz.

2.2. Los corpus lingüísticos

Los corpus son colecciones representativas de textos, delimitadas en función de unos criterios específicos. Desde su aparición en forma digital se ha abogado por su empleo en las clases de lengua extranjera, por ser una fuente de *input* auténtico.

Se han propuesto dos formas de su uso en el aula: la *directa* y *la indirecta* (Pérez Serrano, 2017, p. 77). La forma *indirecta* consiste en que el docente use los datos de corpus en la fase preparatoria de la enseñanza, para extraer textos o ejemplos de uso auténticos (cuya presencia es esencial en los enfoques modernos de enseñanza) y para recabar datos estadísticos; el uso indirecto incluye también consultas realizadas por el propio docente (en especial cuando no es hablante nativo) para aclarar las dudas que pueda tener sobre un fenómeno léxico o gramatical antes de presentarlo en clase. El uso indirecto no se limita a búsquedas realizadas por los docentes; las editoriales

recurren a los corpus para extraer datos auténticos para sus manuales de ELE. Del mismo modo, el uso indirecto implica no solo la extracción de producciones gramaticales, sino también la recopilación de datos negativos, que son esenciales para identificar problemas persistentes de la producción de los alumnos y para adaptar los materiales didácticos en consecuencia (sean estos los materiales de clase creados por los docentes o los manuales preparados por editoriales). Los corpus de aprendientes –por ejemplo, el *Corpus de aprendices de español como lengua extranjera*, *CAES* (Instituto Cervantes, 2014)– son una fuente importante de datos negativos.

El uso *directo* implica el manejo de los corpus por parte de los propios aprendientes tanto en el aula como fuera de él. No obstante, a diferencia de la forma indirecta, el uso directo no ha llegado a integrarse en la práctica cotidiana de la enseñanza. Existen varios trabajos que defienden el uso directo en clase, pero se trata de propuestas teórico-descriptivas (como es el caso de Buyse, 2017; Cruz Piñol, 2012; Elvira-García, 2021). Solo un puñado de trabajos proponen desarrollos prácticos en forma de actividades concretas (entre ellos Álvarez Cavanillas, 2017; Higueras, 2009; Pérez Serrano, 2017), pero ni estas propuestas prácticas lograron introducir la consulta directa como una práctica regular en la enseñanza. El manual de ELE *Frecuencias* (s. f.) de la editorial Edinumen puede considerarse como un avance en este sentido[1]. Dicho manual está enriquecido con textos auténticos extraídos de los corpus del español actual de la RAE (uso indirecto), pero también incluye en el apéndice de su Guía didáctica actividades que ilustran el uso directo de los corpus de la RAE (Edinumen, s. f.). Se trata sin duda de una publicación pionera en este aspecto. Aun así, el hecho de que el uso directo esté limitado al apéndice indica que sigue siendo un recurso marginal y que el uso predominante es el indirecto.

[1] Agradezco esta observación a uno de los evaluadores.

Como se ve, los corpus (especialmente en su uso directo) no se han convertido todavía en herramientas de uso cotidiano en el aprendizaje de lenguas extranjeras pese a sus indudables beneficios y las numerosas publicaciones que abogan por su uso. Consideramos que esta situación puede deberse a dos factores importantes. Por un lado, la mayoría de los textos académicos que exploran su uso en clase son bastante anticuados (se han publicado durante el auge de la lingüística computacional en los años noventa y al principio del nuevo milenio) y operan con sistemas de consulta de esa época, que, desde el punto de vista de hoy, son muy simples. La mayoría de las propuestas propone actividades de consulta de *concordancias*, que son el resultado de una búsqueda en el corpus limitada a una línea de texto (véase Batten, Cornu & Engels, 1989; Gaskell & Cobb, 2004; Kettemann, 1995; Stevens, 1991, entre otros, como ejemplos de uso didáctico de las concordancias). Aunque las concordancias son una función básica y fácil de usar, las posibilidades de su explotación didáctica son muy restringidas: se limitan a la observación de los fragmentos textuales, reflexión sobre su contenido léxico o gramatical y la utilización de las conclusiones del análisis en alguna actividad adyacente.

Por otro lado, aunque los sistemas de consulta han ampliado sus funcionalidades desde los años noventa, siguen siendo herramientas dirigidas principalmente a expertos en la materia. Por lo tanto, presentan ciertas dificultades de consulta para usuarios con poca o ninguna formación lingüística (que es el caso de la mayoría de los aprendientes de ELE). En este sentido, en *CORPES XXI* (*Corpus del Español del Siglo XXI*) y en la versión anotada del *CREA* (*Corpus de Referencia del Español Actual*) de la RAE se ha mejorado sustancialmente la experiencia para el usuario desde la llegada de sus versiones 1.0. Aun así, para realizar una búsqueda simple, el usuario tiene que estar familiarizado, al menos, con los conceptos lingüísticos de *elemento gramatical* y *palabra ortográfica*, *lema* y *forma*. Consideramos que *CORPES XXI* sí tiene potencial didáctico. Para el tipo de contenidos que se abordan este trabajo,

las herramientas especialmente relevantes son *buscar por inventarios* (que permiten ver diferentes formas gramaticales de un lexema) y *coapariciones* (que muestran palabras que se combinan frecuentemente con el elemento buscado, de forma parecida a la de algunas de las herramientas de *Sketch Engine* que describiremos en la sección 3).

Los sistemas modernos de consulta y manejo de corpus, como *Sketch Engine*, con el que trabajaremos en la propuesta, dan un paso más allá. Disponen de una interfaz fácil de manejar, permiten recuperar los resultados de búsqueda en diferentes formatos (textual y gráfico) y disponen de una amplia gama de herramientas de consulta que facilitan el diseño de todo tipo de actividades. Presentaremos *Sketch Engine* con más detalle en la sección 3.

Hay que reconocer que la explotación didáctica incluso de los corpus más modernos presenta varios retos importantes, pero estos son superables en nuestra opinión. Como con cualquier forma nueva de trabajo, se requiere una introducción previa por parte del profesor. Naturalmente, esta introducción debe incluir los aspectos técnicos de su manejo —Boulton (2010, 2012) ha demostrado que para esto es suficiente con una breve introducción—, pero no puede limitarse únicamente a estos aspectos. Según el tipo de corpus y la clase de consulta que se haga, se pueden recuperar demasiados datos o demasiado pocos, y los estudiantes pueden encontrarse incluso con expresiones incorrectas o incompletas (Pérez Serrano, 2017, p. 80). Especialmente con usuarios novatos, la supervisión del profesor en la fase de análisis de datos y formulación de conclusiones es esencial para asegurar que las conclusiones que sacan sean correctas y para que no se desmotiven por la cantidad de datos que no saben interpretar. La capacidad de reflexión es especialmente importante con el uso de los corpus basados en la web, que, al tener muy limitada la capacidad de determinar la procedencia y la autoría de los textos, incluyen producciones de hablantes no nativos y muestras de lenguaje muy descuidado que pueden darse entre hablantes nativos (por ejemplo, en

foros o blogs, donde es posible encontrar muestras donde la gramática y la selección léxica estén algo descuidadas), pero que no es apropiado en la enseñanza del español como segunda lengua. Precisamente por la importancia de este tipo de capacidad reflexiva, consideramos que las actividades con corpus son más recomendables para estudiantes con niveles avanzados del idioma, que disponen de un criterio lingüístico desarrollado y saben interpretar los datos correctamente.

2.3. Los estudiantes avanzados

Se pueden considerar como *estudiantes avanzados* los discentes en la "etapa avanzada-superior o de uso competente de la lengua" (Instituto Cervantes, 2006), que en general corresponde a los niveles C1 y C2 del *Marco común europeo de referencia* (MCER) (Consejo de Europa, 2001). Si nos limitamos al área del léxico, podemos fijar dos criterios esenciales para medir el nivel de semejanza de la expresión discente a la expresión nativa: *la amplitud léxica* (el número de unidades léxicas que conoce el alumno) y la *profundidad léxica* (el conocimiento de las relaciones que existen entre una unidad y otra). A estos dos factores podemos añadir otros dos: la *precisión léxica* (capacidad de diferenciar entre sinónimos gracias al conocimiento de los matices de significado que los distinguen) y la *adecuación léxica*, que se consigue cuando se elige la voz más apropiada distrática y diafásicamente (Capel, 2012, p. 12).

A diferencia de los niveles inferiores (A1-B2), las características del vocabulario de los estudiantes avanzados están determinadas no solo por la instrucción en la LE, sino también por el nivel formativo general y por diversos factores socioculturales (que incluyen los intereses académicos y profesionales). La mayoría de los aprendientes no suelen llegar al nivel de los hablantes nativos, aunque en algunos casos esto puede ocurrir: en áreas específicas relacionadas con los intereses profesionales o académicos específicos del aprendiente, este puede llegar a sobrepasar el conocimiento de un hablante nativo no experto.

El mayor problema de la expresión léxica de estudiantes avanzados es que, incluso cuando alcanzan un tamaño de vocabulario de nivel nativo, frecuentemente conservan rasgos no nativos en su expresión, por lo que siguen *sonando a extranjero*. Esto se debe a un nivel más bajo de la profundidad léxica y un control deficiente de la idiomaticidad. Estas características del conocimiento léxico de estudiantes avanzados se confirmaron, por ejemplo, en el estudio de Bogaards (2001), que reveló que los estudiantes de LE que en algunos casos superaron a sus pares nativos en el tamaño del léxico, se quedaron atrás en lo que se refiere al léxico coloquial, expresiones con carga cultural y el lenguaje formulaico. La asimilación de las reglas que operan la selección léxica nativa (y que determinan, hasta cierto punto, las características de las expresiones idiomáticas) es más difícil de conseguir con los métodos convencionales si la comparamos con el aprendizaje cuantitativo de léxico. Se consigue con la exposición de los estudiantes a la lengua a través de la lectura o el uso de materiales audiovisuales, siempre que el docente haga un esfuerzo adicional para que el aprendizaje no sea pasivo, incidental y, por ende, lento. A este grupo podemos también añadir el trabajo con los corpus, que tiene la ventaja de estimular una reflexión más consciente sobre las propiedades de las unidades léxicas.

2.4. Los corpus propios

Como hemos adelantado, este estudio se centra en el uso de *corpus propios*, que en este trabajo definiremos como 'colecciones de textos ajenos recopiladas por los propios alumnos de LE, quienes las usarán como parte de su aprendizaje'.

Los corpus propios pueden documentar eficientemente el uso de la lengua en un contexto muy acotado. Puesto que el temario de los cursos de idiomas extranjeros casi siempre se compone de temas acotados (correspondientes a las distintas unidades didácticas y campos nocionales), un corpus propio puede aportar al aprendiz datos más relevantes que un corpus general. En un corpus propio, los datos recogidos se pueden acotar en función de

numerosas variables durante el propio proceso de compilación y no *a posteriori*: el formato, el registro, los factores relacionados con su procedencia (autor o autores, lugar de producción y publicación, etc.), a veces incluso su finalidad. Esto permite un mayor nivel de control sobre el tipo de textos que serán recogidos y posteriormente analizados y evita problemas de interpretación de datos (contaminación de datos con producciones no cultas o no nativas a las que hemos aludido en la sección anterior, por ejemplo). Un control de este tipo no sería posible con las herramientas de filtración de los corpus convencionales.

Ya hemos comentado que solo existen unos pocos trabajos dedicados al uso directo de corpus en la enseñanza de segundas lenguas. Ninguno de estos trabajos se centra en los corpus propios. Los estudios que sí lo hacen analizan su uso en otras áreas afines a la enseñanza de lenguas extranjeras, como la traducción o la escritura académica. El artículo de Zhao y Shi (2015), por ejemplo, examina la creación de un corpus paralelo propio para la enseñanza de la traducción. En el estudio presentado en Pérez-Carrasco (2023), se diseñó y se empleó una secuencia didáctica para estudiantes de la traducción científico-técnica. Los alumnos compilaron sus propias colecciones de textos que después usaron en la traducción de textos científico-técnicos. Los trabajos de Tribble y Wingate (2013), Lee y Swales (2006) y Charles (2014) tratan sobre la escritura académica inglesa (*EAP: English for Academic Purposes*). Maia (2005) y Zanettin (2002) experimentaron con el uso de corpus propios en el aula de lenguas extranjeras, pero a diferencia de nuestro estudio no propusieron ninguna actividad concreta centrada en su uso. Los corpus propios en su investigación sirven como una herramienta de apoyo para realizar actividades convencionales, y en su uso se parecen a consultas de los diccionarios. El estudio de Smith (2011, 2020) se parece al nuestro en que se centra en las lenguas extranjeras, hace uso de corpus propios y utiliza *Sketch Engine*. No obstante, no propone actividades concretas. Hasta donde sepamos, la técnica que presentaremos en el siguiente apartado sería la primera de este tipo.

3. PROPUESTA DIDÁCTICA DE USO DE CORPUS PROPIOS

Basándonos en las características de la adquisición del léxico por parte de aprendientes avanzados que hemos revisado brevemente en la sección 2, hemos desarrollado una serie de actividades ejemplares con las que intentamos demostrar el potencial didáctico de nuestra técnica basada en el enfoque léxico y la consulta de los corpus propios. A continuación, presentaremos esta propuesta de una manera general y más bien teórica, centrándonos en las posibilidades que ofrecen los corpus propios y el sistema de consulta y gestión de corpus *Sketch Engine*, del que haremos uso para elaborar las diferentes actividades.

Sketch Engine fue creado en 2003 por la compañía Lexical Computing y se diseñó en colaboración con el equipo de *Natural Language Processing Centre* de la Universidad de Masaryk de Brno. A diferencia de los bancos de datos de la RAE que hemos mencionado en los apartados anteriores (*CREA* o *CORPES XXI*), *Sketch Engine* no es un corpus sino una *herramienta de consulta y gestión de corpus*. Como tal, permite consultar una multitud de corpus monolingües (el *BNC* o los corpus web *TenTen*, entre muchos otros) y corpus paralelos (por ejemplo, *Europarl*) a través de una interfaz de uso fácil y con una amplia gama de opciones de búsqueda. A efectos de este trabajo, es especialmente importante el hecho de que ofrezca la posibilidad de crear corpus propios, basados en textos en línea (con enlaces URL) o en archivos de texto subidos por el usuario. Los corpus propios se pueden compilar en más de 140 lenguas, pero el tipo de anotación automática textual que se ofrece es diferente para cada lengua. Para el español están disponibles todas las funciones importantes, entre ellas las siguientes:

1. Anotación morfosintáctica (*POS tagging*): un etiquetador automático (*tagger*) asigna a cada palabra una categoría gramatical y otros rasgos morfosintácticos –género, número, tiempo verbal, etc.– (Kilgarriff, Rychly, Smrz & Tugwell, 2004, p. 109)

2. Análisis sintáctico (*parsing*): es un proceso realizado por un programa automático (*parser*), que clasifica las unidades léxicas según su función sintáctica en la oración.
3. Lematización: cada palabra gráfica se relaciona con el lema que le corresponde.
4. Compilación automática de *Word Sketches* que resumen el comportamiento contextual de la palabra dada: identifican las unidades léxicas que la acompañan en diferentes posiciones sintácticas (Kilgarriff, Rychly, Smrz & Tugwell, 2004, p. 107).

Al tratarse de herramientas automáticas, no son necesarios conocimientos de lingüística computacional por parte del usuario, pero sí conviene que tenga conocimientos lingüísticos básicos y cierta capacidad de reflexión, porque estos programas en ocasiones se equivocan. En el caso del español, por ejemplo, es común que etiqueten como objeto directo un sujeto pospuesto; también se pueden confundir las categorías gramaticales de palabras homógrafas, por ejemplo. El usuario debería ser capaz de identificar estos fallos cuando analiza los datos.

A diferencia de otros sistemas de consulta, *Sketch Engine* clasifica las coapariciones de palabras (*Word Sketches*) según la función sintáctica que cumplen las palabras recuperadas con respecto al término de búsqueda (por ejemplo, si el término de búsqueda es un verbo, registra los sujetos y los objetos directos que aparecen con él en el contexto). En este sentido, *CORPES XXI* solo las clasifica según su categoría gramatical: por ejemplo, presenta los sustantivos que se combinan con el verbo sin diferenciar su función sintáctica con respecto a ese verbo. Otra ventaja que tiene *Sketch Engine* consiste en la métrica que usa para calcular la fuerza del vínculo en las coapariciones, que se llama *LogDice*. A diferencia de las fórmulas que –como MI score (*Mutual Information*, véase Gablasova, Brezina & McEnery, 2017), usado en *CORPES XXI*– solo tienen en cuenta la frecuencia de coaparición (comparan el número de veces que las palabras aparecen juntas con el número de veces que aparecen

por separado) y por tanto tienden a sobrevalorar algunas combinaciones hápax o erróneas, *LogDice* incluye en el cálculo la relación sintáctica entre las palabras, lo que se traduce en un cómputo más exacto. Volveremos sobre este tema en la sección 3.1.

Antes de usar cualquier herramienta tecnológica hay que introducirla debidamente. Este es el caso de los corpus digitales en general: en los estudios que hemos citado en la sección introductoria (Boulton, 2010, 2012), se hacía una introducción breve (de 5 a 10 minutos) sobre su manejo. *Sketch Engine* es un sistema más complejo que el que usó Boulton, por lo que en su caso la presentación previa debería ser algo más extensa. Como ha advertido uno de los evaluadores, el nivel de especialización de *Sketch Engine* puede ser más alto que el de *CORPES XXI* en al menos algunos aspectos, por ejemplo, en que *Sketch Engine* hace uso del lenguaje CQL (Corpus Query Language) para las búsquedas avanzadas. Por eso recomendamos que las consultas iniciales en *Sketch Engine* en el aula se limiten a las búsquedas categorizadas como básicas, en las que el usuario introduce solamente el término buscado y todos los parámetros adicionales se determinan de forma automática. En *CORPES XXI* no es posible este tipo de búsqueda simplificada. Consideramos que una introducción a las herramientas de *Sketch Engine* en versión básica se podría hacer en una hora, y que se podría segmentar en explicaciones breves de 5-10 minutos sobre cada una de las herramientas, realizadas en diferentes momentos del curso.

Para que las actividades que proponemos en este apartado se puedan llevar a cabo, los corpus propios que se usen deben ser suficientemente grandes, lo que es especialmente importante para las herramientas *Word Sketch, Word Sketch Diferencial* y *N-gramas*. Aunque resulta difícil fijar una cifra mínima por la cantidad de variables que existen, basándonos en las pruebas que hemos hecho, el tamaño mínimo recomendable sería de unas 20 000 palabras para temas muy acotados. Aunque el número parezca grande, gracias a la opción de compilar los corpus usando textos en línea (copiando

sus *URL*), la compilación es relativamente rápida: una estimación realista sería entre 10 y 15 minutos por un corpus temático.

En el siguiente subapartado (3.1.) introduciremos brevemente las herramientas de *Sketch Engine* de las que haremos uso, en la sección 3.2. describiremos el corpus propio de muestra que hemos compilado, y en la última sección (3.3.) presentaremos las actividades agrupadas según la herramienta usada en cada caso.

3.1. Herramientas de Sketch Engine *usadas en la propuesta*

En este apartado describimos las herramientas *Concordancia*, *Word Sketch*, *Word Sketch Differencial* y *N-gramas*[2].

Las concordancias son la función más básica de cualquier herramienta de consulta de corpus. Como hemos explicado en la sección 2.2., las concordancias son fragmentos de texto limitados a una línea, que son el resultado de una búsqueda en el corpus. Tras la definición de los parámetros de búsqueda, se presentan todas las apariciones de la expresión buscada en el formato KWIC (*Key Word in Context*) el elemento buscado aparece en el centro de la línea de texto (véase Figura 1)) o como frases sueltas. En la Figura 1 reproducimos un fragmento de las concordancias para el lema *violencia* a partir del corpus de muestra que presentaremos en el siguiente apartado.

La herramienta *Word Sketch* permite visualizar en forma de listas las palabras que cumplen ciertas funciones sintácticas con respecto al término buscado: sujeto, objeto, modificador, predicado selector, etc. La cantidad y tipo de datos que se muestran depende del tamaño del corpus y la categoría gramatical del término buscado. Por ejemplo, para los verbos y adjetivos se identifican como modificadores los adverbios, y para los sustantivos los adjetivos. La Figura 2 contiene el resumen de Word Sketch generado para el lema *denunciar*.

[2] En este trabajo usaremos la denominación oficial española de las herramientas de *Sketch Engine*. Únicamente adoptaremos nuestra propia traducción para la herramienta *Word Sketch Difference* –a la que nos referiremos como *Word Sketch Diferencial*– porque consideramos que la traducción oficial (*Diferencia Sketch)*, no refleja de manera precisa la funcionalidad de la herramienta.

Figura 1: *Fragmento de concordancias para el lema* violencia

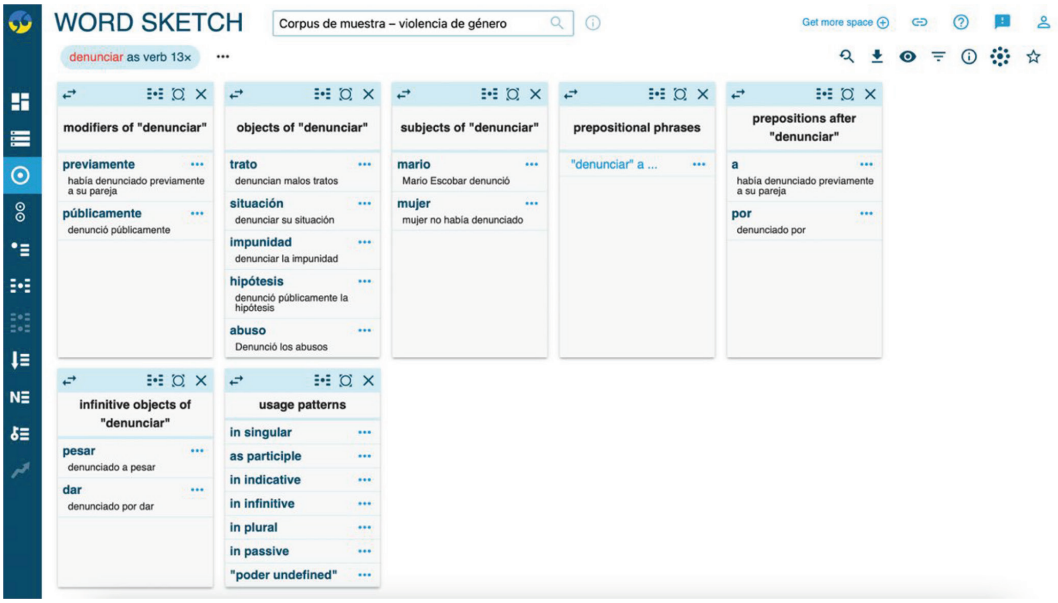


Figura 2: *Resumen de* Word Sketch *para el lema* denunciar

Además de listados de palabras, en *Word Sketch* se pueden ver representaciones gráficas, parecidas a mapas conceptuales, en las que distintas funciones sintácticas están incluidas en zonas marcadas con colores diferentes y donde también están gráficamente representadas la *frecuencia absoluta* de estas

combinaciones (cuántas veces en total aparecen en los datos) y la fuerza del vínculo existente entre dos palabras que coaparecen en el contexto (el colocativo y la base). Esta fuerza se mide a través del indicador *LogDice*, que ya hemos mencionado. *LogDice* compara la frecuencia de coaparición del colocativo y la base unidos por una relación sintáctica específica con la frecuencia de sus combinaciones con otras palabras (con la misma o distinta relación sintáctica). En la representación gráfica de *Word Sketch*, la fuerza de la coaparición está reflejada a través de la distancia entre la base y el colocativo[3]. La frecuencia absoluta se representa por el tamaño del círculo que rodea a los colocativos. La información que aportan *LogDice* y la frecuencia absoluta es muy útil para el aprendizaje del lenguaje formulaico y las expresiones idiomáticas. Con datos de la frecuencia absoluta de una expresión los estudiantes pueden deducir si es una expresión de uso real y si es común. El indicador *LogDice*, por su parte, les informa sobre el grado de fijación sintáctica de una combinación de palabras. En el caso de una expresión fija, el trabajo con corpus puede evitar que el alumno produzca combinaciones léxicas poco adecuadas.

Como ilustración, haremos una consulta de Word Sketch en el corpus web del español *esTenTen* de la palabra *caballo*. Entre las combinaciones con la frecuencia absoluta más alta encontramos, por ejemplo, *caballo negro* –frecuencia absoluta de 2500 casos– (Lexical Computing Ltd., 2021). A pesar de su frecuencia, no es una expresión fija: tanto la base como el colocativo pueden ser sustituidos por otras unidades (por ejemplo, como en *perro negro* o *caballo grande*). Esto también se refleja en la puntuación baja de *LogDice* de *caballo negro*: 5.2 (Lexical Computing Ltd., 2021). Por su parte, la expresión *caballo ganador* que aparece también en los resultados tiene una frecuencia absoluta parecida a la de *caballo negro,* pero su fijación interna es mucho mayor: en este caso, *LogDice* –8.2– (Lexical Computing Ltd., 2021)

[3] En los listados de *Word Sketch*, el *LogDice* (o fuerza de coaparición) se indica de forma directa, a través de un número (hasta 14).

le informa al aprendiente de que se trata de una combinación relativamente estable que no admite mucha variación (no son comunes las combinaciones *perro ganador* o *caballo perdedor*, por ejemplo).

La herramienta *Word Sketch Diferencial* permite comparar los resúmenes de *Word Sketch* de dos palabras para ver en qué se parece su comportamiento contextual y en qué se diferencia. En la Figura 3, reproducimos la comparación de *Word Sketch Diferencial* para los sustantivos *cuerpo* y *cadáver*. La combinatoria de *cuerpo* está marcada en verde, la de *cadáver* en rojo, y en las franjas intermedias aparecen los elementos contextuales que estas palabras tienen en común.

Por último, la herramienta *N-gramas* facilita la búsqueda de expresiones pluriverbales, que a menudo no son composicionales. Estas se presentan en una lista junto con su frecuencia absoluta. La búsqueda se puede acotar por la extensión de la expresión (por el número de palabras que la forman) y otros criterios (como la secuencia de letras que debe contener la expresión). La Figura 4 contiene expresiones estables de tres y cuatro palabras detectadas en el corpus de muestra.

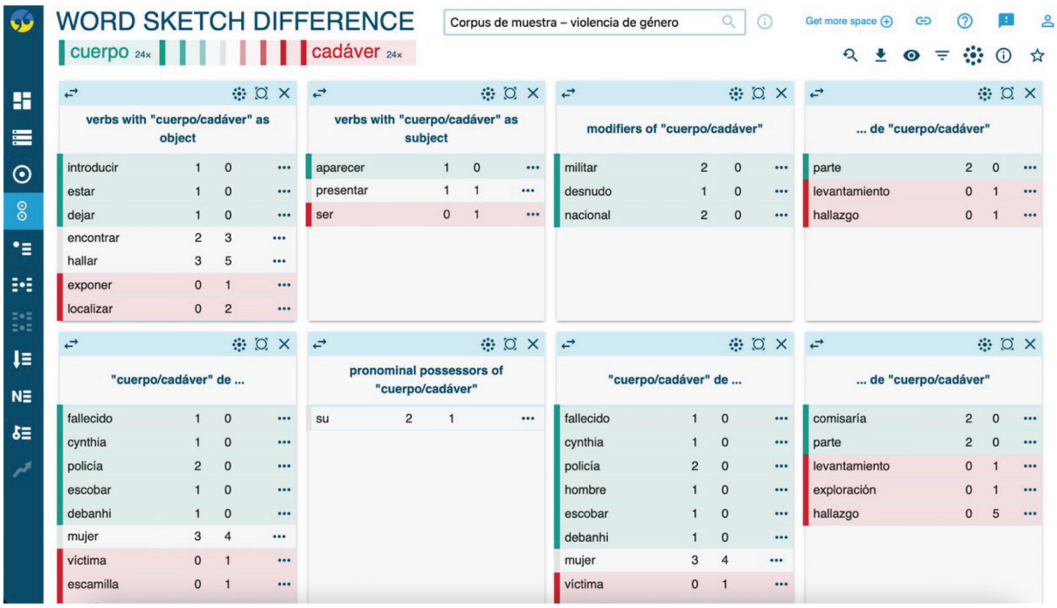
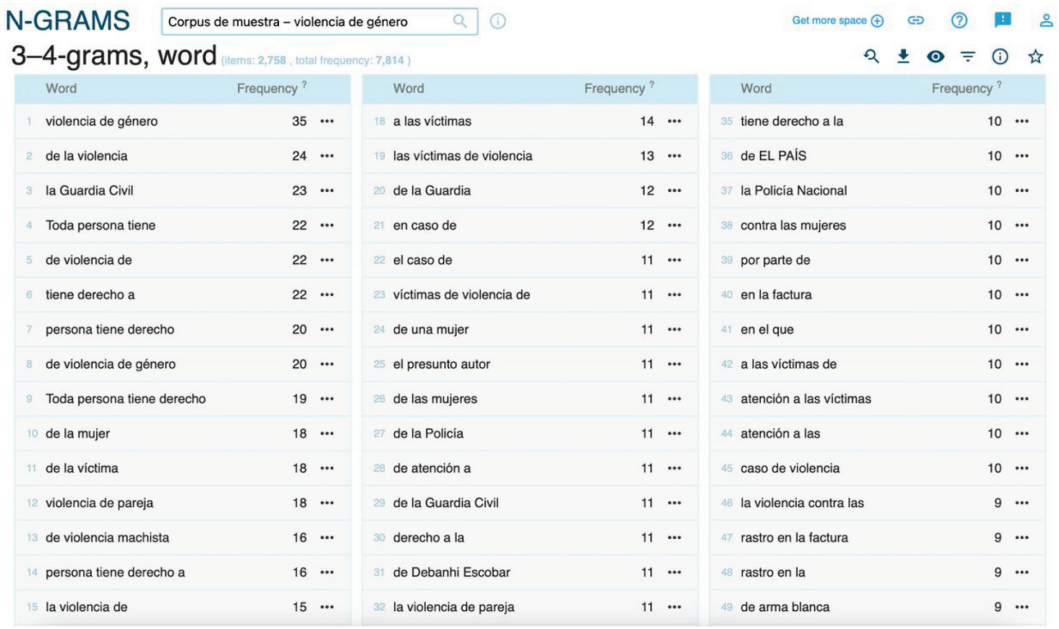


Figura 3: *Comparación de la combinatoria de* cuerpo *y* cadáver *a través de* Word Sketch Diferencial



N-GRAMS Corpus de muestra – violencia de género

Get more space

3–4-grams, word (items: 2,758 , total frequency: 7,814)

	Word	Frequency		Word	Frequency		Word	Frequency
1	violencia de género	35	18	a las víctimas	14	35	tiene derecho a la	10
2	de la violencia	24	19	las víctimas de violencia	13	36	de EL PAÍS	10
3	la Guardia Civil	23	20	de la Guardia	12	37	la Policía Nacional	10
4	Toda persona tiene	22	21	en caso de	12	38	contra las mujeres	10
5	de violencia de	22	22	el caso de	11	39	por parte de	10
6	tiene derecho a	22	23	víctimas de violencia de	11	40	en la factura	10
7	persona tiene derecho	20	24	de una mujer	11	41	en el que	10
8	de violencia de género	20	25	el presunto autor	11	42	a las víctimas de	10
9	Toda persona tiene derecho	19	26	de las mujeres	11	43	atención a las víctimas	10
10	de la mujer	18	27	de la Policía	11	44	atención a las	10
11	de la víctima	18	28	de atención a	11	45	caso de violencia	10
12	violencia de pareja	18	29	de la Guardia Civil	11	46	la violencia contra las	9
13	de violencia machista	16	30	derecho a la	11	47	rastro en la factura	9
14	persona tiene derecho a	16	31	de Debanhi Escobar	11	48	rastro en la	9
15	la violencia de	15	32	la violencia de pareja	11	49	de arma blanca	9

Figura 4: *Expresiones estables de 3 y 4 palabras detectadas por N-gramas*

Todas estas herramientas proporcionan la frecuencia (absoluta y/o relativa) de aparición de los resultados y otros datos estadísticos, indicativos, por ejemplo, de la fuerza de la relación contextual (*LogDice*) a la que nos hemos referido en la descripción de la herramienta *Word Sketch*.

3.2. Corpus de muestra

Para ilustrar la aplicación de corpus propios en la enseñanza de ELE se ha compilado un corpus de muestra a través de *Sketch Engine* (opción "Buscar textos en la web"). Se ha elegido un tema de actualidad que podría tratarse en actividades ELE de nivel avanzado: la violencia de género. Se han recopilado textos de distintas fuentes, entre ellas artículos publicados en *El País, El Mundo, RTVE* y textos informativos del Gobierno de España. El corpus de muestra contiene un total de 30 000 palabras.

La creación de un corpus propio con textos sobre el tema tratado es el paso previo para la realización de las actividades que presentaremos a continuación. Puede ser un corpus diferente para cada alumno, pero parece más recomendable que sea el mismo para todo el grupo, basado en los mismos

archivos de texto o en las mismas páginas web. Así se agiliza la compilación y el profesor mantiene un mayor control sobre el trabajo de los alumnos. Además, *Sketch Engine* permite descargar y compartir un corpus propio, que se puede usar en su versión original por todos los alumnos o ser personalizado (por ejemplo, añadiendo materiales adicionales) en función de las actividades posteriores que se quieran plantear.

3.3. Ejemplos de aplicación didáctica de las herramientas de Sketch Engine

En esta sección presentaremos propuestas concretas de aplicación didáctica de *Sketch Engine*. Como se verá, las tareas diseñadas a veces tratan aspectos relacionados con la organización del léxico. Por este motivo, y para no saturar a los alumnos con información sobre el manejo de los corpus, conviene distribuir el uso de las distintas herramientas entre temas o unidades didácticas diferentes. Para afianzar el aprendizaje de clase, recomendamos integrar este tipo de tareas en el trabajo autónomo de los alumnos (por ejemplo, los deberes de casa) y realizar su seguimiento individual.

3.3.1. Concordancias

Se pueden proponer dos formas de uso de las concordancias, a las que nos referiremos como *uso auxiliar* y *uso orientado*. El *uso auxiliar* consiste en el uso de las concordancias como una herramienta de consulta en actividades de clase no relacionadas con los corpus, como sustituto del diccionario. El corpus es una buena alternativa para los diccionarios monolingües y bilingües porque contiene muchos más datos en la lengua meta, como ejemplos y contextos de uso. Por ejemplo, si un alumno quiere ver con qué preposiciones se combina un verbo de régimen preposicional (p.ej., *depender*), en Google tiene que introducir *depender* (o "*depender* + preposición") y después revisar multitud de resultados para encontrar el dato deseado. En cambio, en las concordancias ve directamente los ejemplos de uso con la preposición. La búsqueda se puede afinar aún más con las opciones avanzadas, como *filter context* (filtrar el contexto) combinado con las etiquetas de categoría sintáctica (*POS context*). Así, se podrían buscar todos los ejemplos en los que

el verbo va seguido de una preposición dentro de un marco de proximidad (a una o más palabras de distancia con respecto al verbo). Del uso auxiliar ya han hablado varios autores (véase la sección 2), por lo que no vamos a dedicarle más espacio.

Las *tareas de uso orientado* se articulan alrededor de las concordancias. Se trataría especialmente de actividades que de alguna forma implican el análisis del léxico en el contexto: ejercicios de rellenar huecos, ejercicios léxico-gramaticales (p.ej., sobre cambios de significado asociados con los verbos *ser* y *estar*), etc. Para dar un ejemplo más concreto, nos detendremos aquí en las actividades de rellenar huecos, en las que se presenta un texto con algunos fragmentos omitidos e incluidos en una lista aparte. Sería recomendable elegir un texto que tenga el mismo tema que el corpus propio, pero que no esté incluido en dicho corpus. Conviene también que las palabras omitidas sean específicas del tema tratado y que los alumnos no las conozcan (es importante que entre ellas haya unidades léxicas pluriverbales). Como primer paso, se pide a los alumnos que elijan palabras correspondientes a cada hueco basándose en el texto de la tarea y, si tienen dudas, pueden recurrir a las concordancias como segundo paso. Tras buscar en concordancias una palabra o unidad léxica de la lista, se presenta al alumno un volumen del *input* mucho mayor. Resulta más fácil deducir el significado de las unidades léxicas desconocidas a partir de múltiples ejemplos de uso para luego poder posicionarlas dentro del texto de la tarea.

La segunda tarea que describimos es de tipo productivo y tiene como objetivo la consolidación de las estrategias de uso de los corpus por parte de los aprendientes. Se pide a los alumnos que redacten un texto de un género específico (por ejemplo, un artículo de opinión) en el que aparezcan determinadas unidades léxicas relacionadas con el tema asignado. Estas unidades léxicas (entre tres y cinco para cada estudiante) se asignan o se distribuyen por sorteo. Para usarlas adecuadamente, los alumnos tendrán que analizar

su uso. Para palabras previamente conocidas a nivel pasivo, es una manera de desarrollar la profundidad léxica, porque a través de las concordancias se aprenden nuevos matices semánticos y de uso.

3.3.2. Word Sketch

Como se ha explicado en la sección introductoria de este bloque, la herramienta *Word Sketch* compila listados de las palabras que coaparecen frecuentemente con la voz buscada. Antes de proponer actividades con *Word Sketch*, el docente debe ofrecer una breve introducción al manejo de esta herramienta. La propuesta que presentamos en esta sección hace uso de los mapas conceptuales para reflejar la combinatoria de las unidades léxicas. Este método de trabajo se promueve dentro del enfoque léxico y está justificado por las bases cognitivas de la organización y el procesamiento del léxico mental, que funciona como un sistema relacional complejo.

Para preparar la actividad que presentaremos a continuación, el docente debe elegir entre tres y cinco palabras (bases) que tengan cierta frecuencia en el corpus propio y que pertenezcan a una de las principales categorías gramaticales léxicas (sustantivos, verbos, adjetivos y adverbios), que son las que tienen mayor carga semántica y suelen tener restricciones de selección claramente definibles. Para hacerlo, el profesor puede usar otra función de *Sketch Engine*, *Word List* (listado de palabras), que extrae por orden de frecuencia absoluta todas las palabras (en nuestro caso, lemas nominales, adjetivales y verbales) que aparecen en el corpus. Para ilustrar nuestra propuesta de actividades, hemos elegido las voces *arrestar, consentimiento, denunciar* y *penal* del corpus de muestra. Esta pequeña lista se da a los alumnos para que analicen su comportamiento contextual con ayuda de *Word Sketch* y posteriormente hagan sus propios mapas conceptuales. Serían parecidos al que se presenta en la Figura 5 para el verbo *denunciar* (las palabras conectadas por líneas azules).

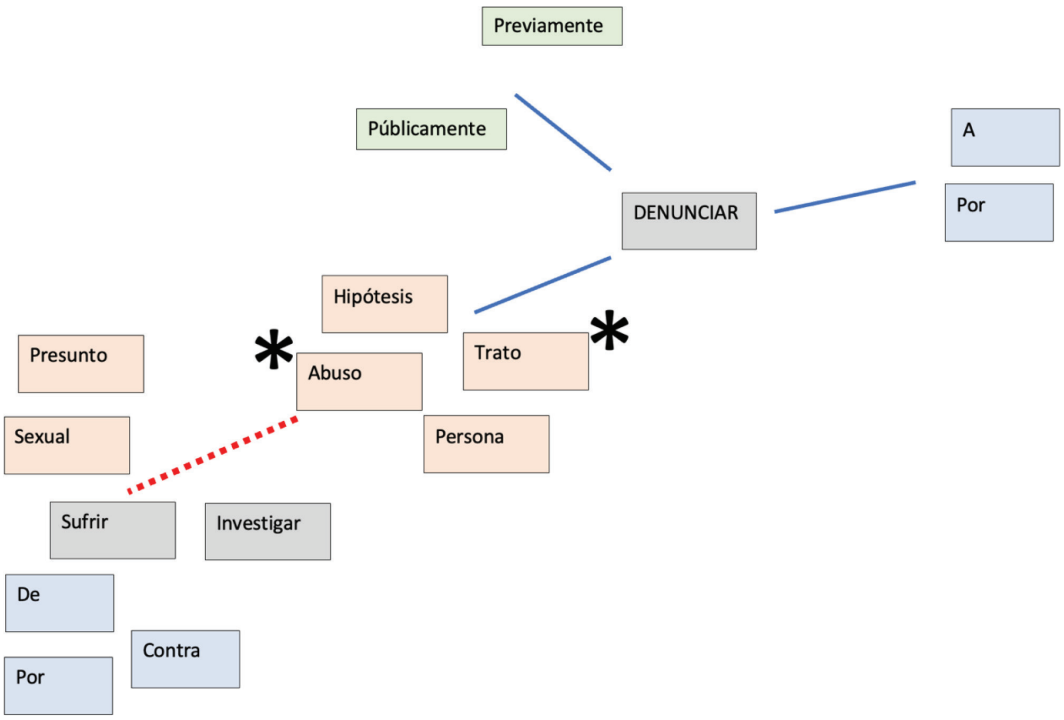


Figura 5: *Mapa conceptual desarrollado basado en el comportamiento contextual del verbo* denunciar

Siguiendo el formato de las visualizaciones gráficas de Word Sketch, que hemos mencionado en la sección 3.1., están marcadas con colores diferentes las palabras que cumplen diferentes funciones sintácticas con respecto a *denunciar*: los modificadores adverbiales (*previamente* y *públicamente*) aparecen en verde, los objetos directos nominales (*hipótesis, trato, abuso* y *persona*) en beis, y las preposiciones *a* y *por* en gris. Una vez diseñados los mapas conceptuales, se puede pedir a los alumnos que los comparen con las visualizaciones gráficas de Word Sketch y que marquen en sus mapas las combinaciones léxicas con niveles elevados de fijación (combinaciones no variables, a las que hemos aludido en la sección 3.1.). En el esquema ejemplar (figura 5), estas combinaciones fuertes se han marcado con el símbolo de asterisco: ambas combinaciones, *denunciar* + *abuso* y *denunciar* + *trato,* disponen de una calificación *LogDice* alta en el corpus de muestra (12,5 y 12,7, respectivamente).

En el siguiente paso de la actividad, se puede proponer a los alumnos expandir los mapas conceptuales añadiendo el contexto de algunas de las palabras incluidas en el mapa. En la Figura 5 ilustramos cómo se puede hacer con *abuso* (objeto directo de *denunciar* de la búsqueda original); la parte añadida se ha marcado con una línea discontinua roja. Además, si el corpus no es muy grande, el mapa se puede enriquecer con unidades léxicas de un corpus general. Así se fomentaría en los alumnos la capacidad de buscar ejemplos propios.

Por último, propondremos una actividad adicional que hace uso de un listado específico de *Word Sketch* (*usage patterns* 'patrones de uso') para los verbos y otras palabras relacionales, que resume su comportamiento gramatical (sus usos en indicativo o subjuntivo, su aparición en estructuras reflexivas, pasivas, etc.). Aunque nuestra propuesta trata de la enseñanza de léxico, este está, en el enfoque léxico, muy estrechamente vinculado con la gramática. En la actividad que proponemos, se trata precisamente de que los alumnos analicen los patrones de uso de la palabra buscada y los relacionen con las modulaciones semánticas que puede experimentar la palabra cuando se usa como parte de estos patrones. Se usarán dos herramientas de forma combinada: *Word Sketch* y las concordancias. En la fase preparatoria, el docente debe elegir entre tres y cinco palabras que presenten algún tipo de dificultad gramatical en su uso, como ser compatibles con el indicativo y el subjuntivo, o con los verbos *ser* y *estar*. Deben ser palabras relacionadas con el tema tratado y suficientemente frecuentes en el corpus. La tarea de los alumnos consiste en explicar las diferencias semánticas asociadas a estas alternancias sintácticas. Para ilustrarlo, tomaremos como ejemplo el verbo *detener* del corpus de muestra. Al consultar el listado *usage patterns*, los alumnos comprobarán que aparece en la construcción <ser + participio> (*ser detenido*) y <estar + participio> (*estar detenido*). Para explicar las diferencias semánticas entre ambas construcciones, tendrán que acceder a las concordancias (pueden hacerlo directamente desde *Word Sketch*) y, basándose en los ejemplos

de uso, llegar a la conclusión de que la pasiva *ser detenido* alude a un acto y la construcción atributiva *estar detenid*o al estado que es el resultado de dicho acto.

3.3.3. Word Sketch Diferencial

La función de la herramienta *Word Sketch Diferencial* es contrastar el comportamiento gramatical de dos palabras. Como hemos visto en la sección 2.3., la precisión léxica es uno de los principales parámetros que definen el dominio de una LE. En niveles avanzados, se logra a través de una variedad de estrategias: actividades de sustitución de palabras comodín por terminología específica, ejercicios de diferenciación semántica de términos cognados y de cuasisinónimos, etc. La actividad que proponemos se centra precisamente en la comparación de sinónimos y cuasisinónimos en función de su comportamiento contextual.

Como con todas las herramientas de *Sketch Engine* mencionadas en esta sección, antes de la actividad el docente debe explicar para qué y cómo se usa *Word Sketch Diferencial*. Para detectar los (cuasi)sinónimos, el profesor puede hacer uso de la herramienta de frecuencia (*Wordlist*), que hemos introducido antes. Elegiría entre tres y cinco pares de sinónimos. En nuestro corpus de muestra, por ejemplo, aparecen las siguientes parejas: *crimen-delito, cuerpo-cadáver, hallar-encontrar* y *morir-fallecer*. Los alumnos tienen que introducir estas palabras en el buscador de *Word Sketch Diferencial* y analizar las diferencias que existen en cada par (preguntas guía en el ejercicio ayudarán a los alumnos enfocar su búsqueda, por ejemplo: *¿Cuál puede ser un crimen y un delito?"*). En el análisis subsecuente se ve, por ejemplo, que los nombres *crimen* y *delito* se usan con modificadores diferentes: *machista* solo se usa con *crimen*, y *grave* preferentemente con *delito*. En el siguiente paso, se puede pedir a los alumnos que confirmen su análisis a través de un corpus de referencia de gran extensión (por ejemplo, *CORPES XXI* o *esTenTen18*).

Como ya hemos dicho, las actividades con *Word Sketch* y *Word Sketch Diferencial* son especialmente recomendables para el aprendizaje del lenguaje formulaico con un nivel alto de fijación: por un lado, contribuyen a ampliar el vocabulario de los aprendientes mediante la introducción de sinónimos y, por otro lado, les ayudan a detectar las diferencias combinatorias que existen entre los sinónimos. Desde la perspectiva de selección libre, no existe ninguna razón por la que no se pueda usar la combinación *delito machista,* por ejemplo: se trata de una combinación semánticamente válida y gramaticalmente correcta (hemos visto un caso parecido con *caballo ganador*). No obstante, no se da en el uso real, como confirman los datos.

3.3.4. N-gramas

La herramienta *N-gramas* detecta expresiones estables de palabras de diferente nivel de fijación y las ordena según su frecuencia de aparición. Puesto que *N-gramas* muestra la frecuencia absoluta de cada expresión detectada, el estudiante puede distinguir entre combinaciones de voces frecuentemente usadas por los hablantes nativos por un lado y los hápax o usos ocasionales (por ejemplo, invenciones de hablantes específicos) por otro lado. La búsqueda en *N-gramas* se puede acotar de varias maneras: por la secuencia de letras que debe contener la expresión, por el número de palabras gráficas, por frecuencia, etc. Si no se acota la búsqueda, se pueden encontrar expresiones previamente desconocidas. Obviamente, una búsqueda no restringida en un corpus general extenso (como *esTenTen18*) devolvería muchísimos resultados, únicamente clasificados por frecuencia. Por ello, su utilidad didáctica sería limitada. En tal caso, sería más eficiente utilizar las concordancias para buscar expresiones ya conocidas por los alumnos o previamente preparadas por el profesor. En cambio, con un corpus temáticamente acotado (como el nuestro) ambos tipos de búsqueda –acotada y sin acotar– serían viables y se podrían aplicar para extraer expresiones específicas de un tema

determinado dependiendo de si se quiere localizar expresiones ya conocidas o todas las expresiones, incluidas las desconocidas.

Como con la herramienta *Word Sketch*, en las dos actividades con *N-gramas* que proponemos aquí se hace uso de mapas conceptuales como método para representar y asimilar las relaciones léxicas, semánticas y asociativas, pero en este caso se pone énfasis en las expresiones pluriverbales (los resultados de búsquedas en *Word Sketch* y *Word Sketch Diferencial* siempre son unidades léxicas compuestas por dos palabras: la base y el colocativo; por lo contrario, *N-gramas* devuelve resultados multipalabra que en la búsqueda en *Word Sketch* no aparecerían). En la primera actividad se hace una búsqueda no restringida de expresiones con *N-gramas* en el corpus temáticamente acotado; los alumnos limitan la extensión de las expresiones a un mínimo de dos palabras y un máximo de cinco. En el caso de nuestro corpus de muestra, se extraerían expresiones como *violencia {de género / de pareja / contra las mujeres}, parte de lesiones, denuncia por violencia de género, presunto autor del crimen, igualdad de oportunidades*, etc. En el siguiente paso, el docente propondría a los alumnos diseñar un mapa conceptual (en este caso semántico) con estas expresiones: por ejemplo, el bloque representado por *violencia {de género / de pareja / contra las mujeres}* estaría relacionado por un vínculo causal con *parte de lesiones* y con un vínculo agentivo con *presunto autor del crimen*.

En la segunda actividad se haría uso de la búsqueda avanzada de *N-gramas*, para acotarla a expresiones que contengan una palabra concreta. Este paso permite extraer grupos acotados de expresiones, que podrían pasar desapercibidas en un listado general. En nuestro corpus, el sustantivo *violencia* aparece en las expresiones *violencia {de género / de pareja / contra las mujeres}*, *víctimas de violencia, caso de violencia, pacto contra la violencia de género*, etc. Los resultados de esta búsqueda se podrían combinar con el mapa conceptual más general, confeccionado en la primera actividad.

A diferencia de *Word Sketch*, *N-gramas* no clasifica los componentes de cada expresión por su función sintáctica. Es algo que se puede proponer como tarea a los alumnos, como un repaso de contenidos gramaticales previamente adquiridos.

4. CONCLUSIONES

Los corpus son una herramienta de potencial indudable para muchas de las ramas de la lingüística, entre ellas la enseñanza de lenguas extranjeras, pero por diversos motivos el auge de la lingüística computacional en los años noventa no propició su aplicación en la práctica docente. El objetivo de este trabajo ha sido precisamente ilustrar este potencial proponiendo una nueva forma de su uso en el aula de idiomas. Aquí nos hemos centrado en los beneficios que ofrece la consulta de corpus cuando se enmarca dentro del enfoque léxico de Lewis. En concreto, hemos argumentado que el trabajo con corpus puede beneficiar especialmente a los estudiantes avanzados, por las características de la adquisición léxica en este nivel de competencia.

Usando diferentes herramientas del sistema de consulta y gestión de corpus *Sketch Engine* (*Concordancia*, *Word Sketch*, *Word Sketch Diferencial* y *N-gramas*), hemos diseñado una muestra de actividades en las que se explota un corpus propio temáticamente acotado. Estas actividades se han centrado principalmente en las expresiones pluriverbales (su detección y sus propiedades sintácticas, entre ellas su grado de fijación), y los sinónimos y cuasisinónimos (su identificación en los textos y sus diferencias de uso dentro de combinaciones libres de palabras y dentro de expresiones pluriverbales). Hemos mostrado cómo se puede explotar la información estadística que proviene de los corpus (la frecuencia absoluta y la métrica de la fuerza de conexión contextual *LogDice*) para introducir en el aula estos y otros fenómenos léxicos, que se trabajan y se asimilan sobre todo en los niveles avanzados de competencia en segundas lenguas.

La eficacia de la técnica propuesta y de las actividades diseñadas habrá de ser confirmada en un estudio empírico amplio. El mismo estudio contribuirá a verificar la adecuación de los parámetros más técnicos, que aquí hemos fijado de forma tentativa, entre ellos el tamaño del corpus autocompilado y la forma de recolección de los datos. En este estudio, las expresiones pluriverbales se abordarán como una de las manifestaciones particulares de la combinatoria léxica, cuyo tratamiento en el aula de segundas lenguas no se suele abordar de forma sistemática a pesar de su importancia para poder alcanzar niveles altos de competencia léxica, sintáctica y discursiva.

FINANCIACIÓN

La publicación es parte del proyecto "Corpus de composicionalidad e informatividad léxica: anotación, análisis y aplicaciones (INFOLEXIS)" (IP Olga Batiukova, referencia PID2022-138135NB-I00), financiada por MCIN/AEI/10.13039/501100011033 y por el FSE+. Su realización ha sido financiada por un contrato predoctoral asociado a este proyecto.

La publicación es parte de la ayuda PREP2022-000882, financiada por MCIN/AEI/10.13039/501100011033 y por el FSE+.

REFERENCIAS BIBLIOGRÁFICAS

Álvarez Cavanillas, J. (2017). Un enfoque léxico en los manuales ELE. En F. Herrera (ed.), *Enseñar léxico en el aula de español: el poder de las palabras* (pp. 57-69). Difusión.

Batten, L., Cornu, A. & Engels, L. V. (1989). The use of concordances in vocabulary acquisition. En C. Laurent & M. Nordman (eds.), *Special language: from humans thinking to thinking machines* (pp. 452-467). Multilingual Matters.

Bogaards, P. (2001). Lexical units and the learning of foreign language vocabulary. *Studies in Second Language Acquisition*, *23*(3), 321-343. https://doi.org/10.1017/S0272263101003011

Boulton, A. (2010). Data-driven learning: taking the computer out of the equation. *Language Learning*, *60*(3), 534-572. https://doi.org/10.1111/j.1467-9922.2010.00566.x

Boulton, A. (2012). Hands-on / hands-off: alternative approaches to data-driven learning. En J. Thomas & A. Boulton (eds.), *Input, process and product: developments in teaching and language corpora* (pp. 153-169). Masaryk University Press.

Buyse, K. (2017). Los corpus como herramientas de aprendizaje de léxico. En F. Herrera (ed.), *Enseñar léxico en el aula de español: el poder de las palabras* (pp. 121-140). Difusión.

Capel, A. (2012). Completing the English vocabulary profile: C1 and C2 vocabulary. *English Profile Journal*, *3*(1), 1-14. https://doi.org/10.1017/S2041536212000013

Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes*, *35*, 30-40. https://doi.org/10.1016/j.esp.2013.11.004

Consejo de Europa (2001). *Marco Común Europeo de Referencia para las lenguas*. División de Lenguas Modernas.

Cruz Piñol, M. (2012). *Lingüística de corpus y enseñanza del español como 2-L*. Editorial Arco Libros.

Edinumen. (s.f.). *Frecuencias*. Edinumen. https://edinumen.es/materiales/frecuencias/

Elvira-García, W. (ed.). (2021). *El uso de corpus en clase de ELE: la lengua real como modelo*. Difusión.

Gablasova, D., Brezina, V. & McEnery, T. (2017). Collocations in corpus-based language learning research: identifying, comparing, and interpreting the evidence. *Language Learning*, *67*(1), 155-179. https://doi.org/10.1111/lang.12225

Gaskell, D. & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System, 32*(3), 301-319. https://doi.org/10.1016/j.system.2004.04.001

Higueras, M. (2009). Aprender y enseñar léxico. *MarcoELE*, *9*, 111-26.

Instituto Cervantes (2006). *Plan curricular del Instituto Cervantes*. Instituto Cervantes. https://n9.cl/ysfzp

Instituto Cervantes (2014). Corpus de aprendices del español: CAES (Versión 2.1.) [corpus]. Instituto Cervantes. https://galvan.usc.es/caes/

Kettemann, B. (1995). On the use of concordancing in ELT. *AAA: Arbeiten aus Anglistik und Amerikanistik*, *20*(1), 29-41.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. En G. Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress* (pp. 105-15). Université de Bretagne-Sud, Faculté des lettres et des sciences humaines. https://n9.cl/gg423

Lee, D. & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: moving from available specialized corpora to self-compiled corpora. *ScienceDirect*, *25*, 56-75. https://doi.org/10.1016/j.esp.2005.02.010

Lewis, M. (2008). *Implementing the Lexical Approach: putting theory into practice*. Heinle.

Lexical Computing Ltd. (2021). esTenTen (estenten18_fl5_1) [corpus]. Lexical Computing Ltd. https://www.sketchengine.eu

Maia, B. (2005). Terminology and translation–bringing research and professional training together through technology. *Meta: Journal des traducteurs. 50*(4), s. p. https://doi.org/10.7202/019921ar

Pérez-Carrasco, M. (2023). The nuts and bolts of corpora: secuencia didáctica para el empleo de corpus en la enseñanza de la traducción científico-técnica. *Hikma: Estudios de Traducción, 22*(1), 307-337. https://doi.org/10.21071/hikma.v22i1.15374

Pérez Serrano, M. (2017). *La enseñanza-aprendizaje del vocabulario en ELE desde los enfoques léxicos*. Arco Libros-La Muralla.

Real Academia Española. (s.f.). *Corpus del español del siglo XXI: CORPES XXI* (Versión 1.1.) [corpus]. Real Academia Española. http://www.rae.es

Real Academia Española. (s.f.). Corpus de referencia del español actual: CREA (Versión 1.0.) [corpus]. Real Academia Española. http://www.rae.es

Rufat, A. & Jiménez Calderón, F. (2017). Aplicaciones de enfoques léxicos a la enseñanza comunicativa. En F. Herrera (ed.), *Enseñar léxico en el aula de español: el poder de las palabras* (pp. 47-55). Difusión.

Sinclair, J. (1991). *Corpus, concordance, collocaction*. Oxford University Press.

Smith, S. (2011). Learner construction of corpora for general English in Taiwan. *Computer Assisted Language Learning*, *24*(4), 291-316. https://doi.org/10.1080/09588221.2011.557024

Smith, S. (2020). DIY corpora for Accounting & Finance vocabulary learning. *English for Specific Purposes*, *57*, 1-12. https://doi.org/10.1016/j.esp.2019.08.002

Stevens, V. (1991). Concordance-based vocabulary exercises: a viable alternative to gap-fillers. *Classroom Concordancing: English Language Research Journal*, *4*, 47-63.

Tribble, C. & Wingate, U. (2013). From text to corpus: a genre-based approach to academic literacy instruction. *ScienceDirect*, *41*, 307-21. https://doi.org/10.1016/j.system.2013.03.001

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press. https://doi.org/10.1017/CBO9780511519772

Zanettin, F. (2002). DIY corpora: the WWW and the translator. En B. Maia, J. Haller & M. Urlrych (eds.), *Training the language services provider for the new millennium* (pp. 239-248). Universidad de Porto.

Zhao, Y. & Shi, J. (2015). Self-compiled on-line parallel corpus in translation teaching. En W. Liu (ed.), *Conference on Education and Teaching in Colleges and Universities* (pp. 68-71). Atlantis Press. https://doi.org/10.2991/cetcu-15.2016.20