ONOMÁZEIN
Revista de lingüística, filología y traducción

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE LETRAS

Léxico mejorado del español para los niveles del MCER A1, A2 y B1, y nociones metalingüísticas de base

Enhanced Spanish lexicon for the CEFR levels A1, A2 and B1, and basic metalinguistic notions

Xavier Blanco Escoda
Universitat Autònoma de Barcelona
España

Número especial – XIV –

Variación(es), enseñanza y traducción: investigación(es) en fraseología

2024

ONOMÁZEIN | Número especial XIV
Variación(es), enseñanza y traducción: investigación(es) en fraseología: 150-167
DOI: 10.7764/onomazein.ne14.08
ISSN: 0718-5758

Xavier Blanco Escoda: Departamento de Filología Francesa y Románica, Facultad de Letras, Universitat Autònoma de Barcelona, España. Orcid: 0000-0001-8210-3668. | E-mail: xavier.blanco@uab.cat

Resumen

Este artículo presenta la metodología de elaboración de un léxico del español (hasta nivel B1 inclusive) en el marco del proyecto europeo iRead4Skills. Se introducen las nociones metalingüísticas de base que se deberían presentar a los discentes a fin de que dicho recurso pueda ser explotado de la manera más adecuada posible. Se presentan y discuten los principales campos de microestructura (en particular, los correspondientes a rasgos sintáctico-semánticos, clases semánticas y ámbitos de especialidad) introducidos en la base de datos. También se examina de manera específica el tratamiento de la fraseología y de la diasistemática.

Palabras clave: léxico; niveles MCER para el español; clases semánticas; iRead4Skills.

Abstract

This paper presents the methodology for developing a lexicon of Spanish (up to and including level B1) within the framework of the European iRead4Skills project. The basic metalinguistic notions that should be presented to students are introduced so that this resource could be exploited in the most appropriate way possible. The main microstructure fields (in particular, those corresponding to syntactic-semantic features, semantic classes and domains) introduced in the database are presented and discussed. The treatment of phraseology and diasystematics is also specifically examined.

Keywords: lexicon; Spanish CEFR levels; semantic classes; iRead4Skills.

1. Introducción[1]

A Robert Gallison, pionero de la didáctica del francés como lengua extranjera, le gustaba repetir que la *vaca* "fundamental" no podía *mugir* y que tampoco era posible *ordeñarla*. Con ello quería decir que la lista del *Français fondamental* (aproximadamente un millar de voces) (Gougenheim y otros, 1956), creada en la década de los 50 del siglo pasado como recurso para la enseñanza del francés, contenía un importante número de entradas prácticamente aisladas, ya que, dado lo limitado del inventario, no era posible formar con ellas frases no triviales (más allá de *Veo una vaca* o *Es una vaca grande*...). Cuando, en el marco del proyecto europeo iRead4Skills (cf. nota 1), emprendimos la compilación de un léxico para el español, así como el establecimiento de una serie de criterios para su organización en niveles de complejidad, decidimos tener bien presentes las palabras del maestro y hacer cuanto estuviese en nuestra mano para que el material léxico presentado no quedase reducido a una lista, sino que formase un todo orgánico a partir del cual se pudiesen desarrollar las competencias lingüísticas básicas. El objetivo del presente artículo es presentar dicho léxico y las nociones metalingüísticas fundamentales en las que se basa.

En lo tocante a iRead4Skills[2], no nos es posible presentar aquí, dadas las limitaciones de espacio, un panorama siquiera sucinto de dicho proyecto. El lector interesado hallará información actualizada en la página web https://iread4skills.com. Digamos, únicamente, que uno de sus objetivos primordiales consiste en evaluar las deficiencias respecto a la competencia lectora que puedan presentarse en adultos, a fin de contribuir a reducirlas mediante un sistema de lectura inteligente que evalúe la complejidad de los textos y sugiera lecturas apropiadas según el nivel del usuario. Asimismo, el proyecto persigue apoyar la adopción y difusión de innovaciones educativas que, por una parte, permitan a los adultos con déficit de competencia lectora la adquisición y el mantenimiento de habilidades transversales duraderas y, por otra parte, faciliten a los formadores y creadores de contenido identificar y adaptar materiales con el nivel adecuado de complejidad para su público-objetivo.

1 La investigación presentada en este artículo ha sido financiada por la Unión Europea en el marco del proyecto iRead4Skills (*Intelligent Reading Improvement System for Fundamental and Transversal Skills Development*) (Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07), coordinado por la Universidade Nova de Lisboa.

2 Es un placer agradecer la colaboración, en diferentes momentos de la investigación presentada, de los siguientes investigadores (vinculados a la UAB, salvo indicación contraria): Roser Gauchola, Keran Mu, Àngels Catena, Lorraine Baqué, Julio Murillo, Marcos García (Universidade de Santiago de Compostela) y Sara Rodríguez (Universidade de Santiago de Compostela). Agradecemos también la colaboración de Raquel Amaro (coordinadora general del proyecto) y de Susana Correia (responsable del Workpackage n.º 3, *Complexity classification and data*) (ambas de la Universidade Nova de Lisboa).

2. Un léxico del español para los niveles A1, A2 y B1

Todo docente acostumbrado a enseñar en niveles básicos sabe qué léxico introducir en función de su público-objetivo y cuándo hacerlo. Dispone también, para el español, de inventarios accesibles en mayor o menor medida (desde productos comerciales hasta descripciones detalladas y rigurosas elaboradas por actores mayores en el ámbito del ELE, Español como Lengua Extranjera). Baste con citar aquí el *Plan Curricular del Instituto Cervantes*, que precisa y presenta los niveles de referencia para el español según las recomendaciones del Consejo de Europa en su *Marco Común Europeo de Referencia* (MCER) para las lenguas, cf. Consejo de Europa (2002) y Centro Virtual Cervantes[3].

No cabe duda de que este último recurso constituye la referencia fundamental en el ámbito del ELE[4]. Ahora bien, dentro del marco de los trabajos para el español en iRead4Skills, nos ha parecido importante disponer de un recurso propio. El grupo de investigación en fonética, lexicología y semántica (*flexsem*) de la UAB ha venido desarrollando, desde los años 90 del siglo pasado, un sistema de diccionarios electrónicos para el español (Blanco, 2001 y 2010) que nos han servido de punto de partida para el establecimiento del léxico mejorado del español para los niveles A1, A2 y B1 al que nos referiremos en este artículo. Utilizamos "mejorado" no por referencia a ningún otro léxico existente, sino únicamente en el sentido de que lo hemos dotado de una microestructura (en desarrollo), cuyas principales categorías presentamos a continuación y que pretenden convertirlo en un recurso dinámico para docentes y discentes.

Empecemos con unas nociones básicas previas[5]. El vocabulario de una lengua está organizado en unidades elementales, que son las unidades léxicas[6]. Una unidad léxica es, bien

3 Cf. https://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular.

4 Respecto a las otras dos lenguas abordadas en el proyecto (portugués y francés), pueden consultarse los trabajos del Centro de Linguística de la Universidade Nova de Lisboa (CLUNL), https://clunl.fcsh.unl.pt/, y del Centre de traitement automatique du langage (CENTAL) (Université Catholique de Louvain), https://uclouvain.be/fr/instituts-recherche/ilc/cental.

5 Las nociones presentadas se inscriben, por una parte, en la tradición de estudios lingüísticos del léxico-gramática extendido (Gross, 2013) y, por otra parte, en el marco de la lexicología explicativa y combinatoria (Polguère, 2016; Mel'čuk y Polguère, 2007). No nos es posible, en el marco de este artículo, definir todos los términos empleados, pero nos parece importante mencionarlos para que el lector tenga al menos una idea general de la panoplia de nociones metalingüísticas que se requieren para el buen aprovechamiento de un recurso lexicográfico.

6 Las unidades subléxicas (morfemas productivos) pueden presentarse por confrontación entre unidades léxicas que se opongan, por ejemplo: *feliz*, *infeliz*. Hagamos notar que no siempre la unidad léxica morfológicamente más simple es la más común, ni la más fácilmente comprensible: *increíble* es mucho más común que *creíble*, *inmediato* que *mediato* e *infinito* que *finito*; *descansar* resulta más frecuente que *cansar*. Es preciso, además, hacer notar al discente que

una "palabra", bien una "expresión fija" tomada en un significado determinado y provista de una determinada forma y una determinada combinatoria. Las nociones "palabra" y "expresión fija" las tomamos como indefinibles operativos (y las explicamos a los discentes exclusivamente por ostensión, mediante ejemplos: *manzana*, *hablar*, *bonito* son palabras; *sin embargo*, *fuegos artificiales* y *dar la mano* son expresiones fijas).

El significado es comprensible y representable (con mayor o menor precisión) mediante la propia lengua (una definición, un conjunto de sinónimos o parasinónimos...) o mediante otros sistemas semióticos (una ilustración, una fotografía, un índice —entendido como tipo de símbolo— o una demostración...). La forma es perceptible (en el caso que nos ocupa, se trata de grafemas). La combinatoria está formada por un conjunto de propiedades de la unidad léxica que no son directamente deducibles ni de su significado ni de su forma, pero que son necesarias para la utilización efectiva de la unidad: su género gramatical; sus colocativos más comunes (para *café*: *café solo*, *café con leche*, *café descafeinado*...); su flexión, en caso de tenerla (forma femenina, plural, tiempos y modos verbales...); su rección, en caso de tenerla (preposiciones regidas, número de complementos de un verbo, etc.).

Cada unidad léxica supone un paradigma de formas flexivas (que puede limitarse a una sola forma). Dos o más unidades léxicas que compartan la misma forma (y, muy a menudo, el mismo paradigma) pueden constituir un vocablo si comparten un vínculo semántico no trivial (así, por ejemplo, *avería*[1] 1 *acontecimiento que provoca un daño en el funcionamiento de un aparato* y *avería*[1] 2 *estado defectuoso que presenta una mercancía* son dos lexemas del mismo vocablo, pero *avería*[2] (*lugar donde se crían aves*) constituye otro vocablo).

El diccionario electrónico de formas simples de *flexsem* (Blanco, 2001) está estructurado en lexemas, es decir, cada entrada corresponde a una unidad léxica univerbal y puede, por tanto, haber múltiples entradas para cada forma, siempre y cuando estas correspondan a distintas unidades léxicas. Las indicaciones semánticas de microestructura permiten distinguir entre distintos lexemas.

Para proceder a la obtención de un léxico para el español (A1-A2-B1), se partió del diccionario mencionado (70 000 entradas). Dos expertos etiquetaron cada entrada con un código según el sistema de representación y plausibilidad propuesto por Mylène Garrigues (1992) para obtener una jerarquización del léxico en cuatro niveles[7]: R0 (unidad sobre la que no

existe una gran cantidad de frasemas morfológicos: *desayunar*, *desgracia*, *submarino*, *bikini*... no son el resultado estricto de la composición de los significados de los morfemas que integran.

7 Presentamos una breve caracterización del método en Blanco (2001: 59). Como explicamos a continuación, para la compilación de este nuevo léxico hemos juzgado necesario repetir en 2024 la asignación de etiquetas de plausibilidad. Se trata de una labor interesante desde el punto

se puede emitir un juicio de plausibilidad), R1P1 (unidad que el experto juzga conocida y de aparición probable en textos), R1P2 (unidad que se juzga conocida, pero de aparición poco probable), R1P3 (unidad que se juzga desconocida o de muy baja probabilidad de aparición).

Una vez consensuados los casos de divergencia entre ambos etiquetadores, el nivel de máxima plausibilidad (R1P1) quedó asignado a un conjunto de casi 15 000 entradas. Puede considerarse que estas representan la competencia media de un hablante nativo del español peninsular con estudios superiores (nivel C2). Este subconjunto fue sometido a una nueva clasificación manual en la que se pidió a los etiquetadores que seleccionasen únicamente aquellas voces consideradas indispensables para comprender textos de nivel umbral (denominamos "umbral" al nivel B1). Se obtuvo, así, una lista de aproximadamente 3 000 voces. Finalmente, se clasificaron estas voces en tres niveles. En esta última fase, intervinieron cuatro locutores, dos de nivel C2 y dos de nivel C1. Se añadió, además, un quinto etiquetado, realizado por una locutora de nivel B1.

Se obtuvo, así, un léxico de 3 032 entradas, de las cuales 608 corresponderían al nivel A1, 1 075 al nivel A2 y 1 350 al nivel B1[8]. Para abreviar, aludiremos de ahora en adelante a nuestro léxico como *lex_umb* (léxico umbral). El método que hemos utilizado no es ni fácilmente reproductible ni escalable, pero nos ha proporcionado una base empírica a partir de la cual precisar el vocabulario que queremos aislar. En estos momentos (enero de 2024), se está acabando de compilar, en el marco del proyecto iRead4skills, un corpus para el español, formado por 200 textos de niveles A1, A2 y B1 (aproximadamente un millón y medio de palabras), que permitirá contrastar el grado de cobertura obtenido y afinar la selección de voces.

3. Rasgos sintáctico-semánticos, clases semánticas y ámbitos de especialidad

Una vez delimitado el *lex_umb*, es fundamental dotar a los discentes de una lengua (ya sea propia o extranjera) de un conjunto de nociones metalingüísticas elementales, ya que estas les serán de extraordinaria utilidad en su proceso de apropiación de la lengua, les proporcionarán una visión mucho más clara de su objeto de estudio y constituirán una base sólida sobre la que desarrollar sus competencias, fundamentalmente, en lo que nos concierne, el vocabulario pasivo y la competencia lectora. Asimismo, estas nociones permitirán tanto

de vista lexicológico (que nos ha permitido, por ejemplo, constatar la importante variación diacrónica que se ha producido en un período de poco más de veinte años), pero extremadamente costosa en términos de tiempo.

8 En fecha de 17 de enero de 2024. Las etiquetas están sujetas a otros procedimientos de revisión (verificación en manuales de aprendizaje, cotejo con corpus) y las cantidades pueden ir variando, aunque no sustancialmente.

a los discentes como a sus docentes evaluar los progresos efectuados y detectar las carencias que subsistan o vayan manifestándose. En este apartado presentaremos algunos instrumentos que deben conocerse en relación con la noción de unidad léxica (recordemos que cada entrada del vocabulario[9] compilado representa una unidad léxica univerbal o lexema). Estas nociones deben ser expresadas mediante campos de microestructura y son cruciales para convertir el glosario descrito en un recurso educativo dinámico, lo que hemos llamado "mejorado".

3.1. Los rasgos sintáctico-semánticos

La primera noción clave corresponde a los rasgos sintáctico-semánticos. En primer lugar, es preciso hacer notar al discente que las unidades léxicas se organizan en categorías gramaticales o clases de palabras y mostrarle (si no la conoce ya) la diferencia entre un sustantivo, un adjetivo, un verbo, un adverbio, una preposición, una conjunción, un pronombre y un artículo. Poco importa si deseamos hacer variar esta tipología y subsumir pronombres bajo nombres, reagrupar estos últimos con los adjetivos o agrupar preposiciones y conjunciones bajo la categoría de adverbios (cosa poco común pero plenamente justificable). Basta con presentar y razonar una clasificación operativa que retome las principales categorías comúnmente aceptadas.

Haremos observar después a los discentes que, entre los sustantivos, pueden distinguirse siete grandes clases, que denominaremos rasgos sintáctico-semánticos:

- nombres de 'humanos' (p. ej. *papá*, *profesor*...)[10];
- nombres de 'animales' (p. ej. *gato*, *pájaro*...);
- nombres de 'vegetales' (p. ej. *árbol*, *rosa*...);
- nombres de inanimados concretos (p. ej. *vaso*, *piedra*...);
- nombres de 'lugar' o locativos (p. ej. *región*, *esquina*...);
- nombres de 'tiempo' o temporales (p. ej. *semana*, *Navidad*...);
- nombres abstractos, entre los que distinguiremos, al menos, entre 'acciones' (p. ej. *paso*, *respuesta*...), 'estados' (*gripe*, *tristeza*...) y 'acontecimientos' (*lluvia*, *accidente*...).

9 Obsérvese que, para referirnos a *lex_umb*, utilizamos indistintamente *léxico* (en su acepción metalingüística), *vocabulario*, *diccionario* o *glosario*. No podemos entrar aquí en las diferencias que presentan estos términos que, de momento, no son relevantes para nuestro propósito.

10 Utilizamos las comillas simples para indicar que nos estamos refiriendo al significado de la unidad léxica o del sintagma que empleamos en cada caso.

El interés de esta división consiste en hacer notar que induce una partición en los predicados que seleccionan cada uno de estos rasgos (cf. 3.2). La gran mayoría de verbos seleccionan, o cuanto menos admiten, un sujeto humano y, por tanto, la clave de acceso a su significado habrá que buscarla en los complementos; por otra parte, los verbos que seleccionen como sujeto alguno de los otros rasgos resultarán más sencillos de caracterizar.

3.2. Clases semánticas

Empezando por los sustantivos humanos, haremos notar que una veintena de nombres del *lex_umb* admiten predicados[11] como *trabajar como N*, *salario de N*, *despedir a N*, *estudiar para N*...[12] y que cuatro de ellos (*alcalde*, *obispo*...), sin llegar a rechazar completamente esta combinatoria, le añaden *nombrar N*, *ser elegido N*. Tenemos, pues, nombres de <profesión> y de <cargo>. Esta sintaxis local, formada por los predicados llamados "apropiados" a <profesión>, es muy distinta de la de otro grupo de humanos como *papá*, *mamá*, *abuelo*, *sobrino*... que están etiquetados como <familiar> y de los que puede decirse *tener un N*, *ser el N de*, *mi N*, como sucede con los nombres de <profesión>, pero no son aceptables *#contrato de <familiar>*, *#despido de <familiar>*[13] o *#cese de <familiar>*.

Haremos observar, igualmente, que un pájaro puede *cantar*, *volar*, *comer* o *morir* y que se puede decir del mismo que tiene *pico*, *alas* y *plumas*[14]. Que cualquier sustantivo que pueda aparecer y que comparta una combinatoria similar (ya forme parte del vocabulario básico, como *paloma*, o no, como *gorrión*) puede ser subsumido bajo la condición de <pájaro>, que constituirá su género próximo. Y que una frase como *Se escucha el gorjeo de un gorrión* puede plantear dificultades de comprensión, pero contiene en sí misma elementos de respuesta que vienen dados por la selección apropiada del verbo *escuchar* (se escucha

11 A partir de combinatorias como las que esbozamos aquí, se introducirán las importantes nociones de predicado, argumento semántico (Gross, 2013: 37) y actualización (Gross, 2013: 183). Limitémonos aquí a un ejemplo caricaturalmente sencillo: en una frase como *Juan se comió una manzana*, *comer* corresponde al predicado; *Juan* y *manzana*, a los argumentos semánticos; las marcas temporales y modales del verbo, el determinante *una* y la forma morfológica de los sustantivos corresponden a la actualización (predicados gramaticales) del esquema predicativo *comer* (*Juan*, *manzana*).

12 Empleamos aquí una forma muy simplificada de representar los predicados y su combinatoria (esquemas proposicionales).

13 Utilizamos el símbolo # para marcar las secuencias asemánticas, sintagmas o frases que, aunque gramaticalmente estén bien formadas, no resultan aceptables en español en condiciones de interpretación normales.

14 Nótese que cada uno de los verbos mencionados (que forman parte de *lex_umb*) no permite aislar por separado la clase <pájaro>, pero la intersección de los cuatro sí arroja grandes posibilidades de caracterizar como tal el objeto seleccionado. Lo mismo sucede con los tres merónimos de ejemplo, aunque estos son más específicos a la clase en cuestión.

un <sonido>) y de su sujeto semántico 'gorrión_pájaro' que presenta una gama limitada de posibilidades vocales comparado con un humano. El opaco *gorjear* puede, pues, ser en gran medida elucidado ('gorjear' ≈ 'cantar').

Mostraremos que un árbol puede *crecer* y *morir*, como un humano y un animal, pero que también se puede *plantar*, *regar* o *cortar*; que 'cortar un árbol' puede denominarse *talar* y que el árbol posee *ramas*. *Pino* o *abeto* no están dentro de las 3 000 voces que conforman nuestro *lex_umb*, pero podemos señalar que es posible *plantar* o *cortar* un 'pino' o un 'abeto', que la frase *talar un pino* es, por tanto, aceptable (pero no *#talar una rosa* o *#talar una mesa*) y que puede decirse igualmente de un pino que *crece* y que tiene *ramas*. El discente podrá, así, incorporar *pino* y el inevitable *abeto* navideño como nombres de <árbol> a su vocabulario B2. Y es interesante observar que, una vez que se haya introducido cierto número de sustantivos que designen árboles, estos permitirán detectar nuevos predicados apropiados, observando que los árboles, como algunos otros vegetales, pero no todos, y nunca un humano ni un animal, se pueden *podar*, *trasplantar*, *repoblar* o *injertar*.

Algunos inanimados concretos serán apropiados en contextos como *vestir un N*, *probarse un N*, *talla de N*, *N estrecho*, *N a rayas*. Se trata de la clase <prenda de vestir>, que está representada por unos 40 nombres[15] de nuestra lista: *abrigo*, *camisa*, *pantalón*... Muy distinta será la combinatoria de otros 40 N que corresponden a <partes del cuerpo> (*barriga*, *boca*, *brazo*...).

Entre los abstractos, hallaremos un grupo de unos 30 sustantivos[16] apropiados en combinatorias como *sentir N*, *producir N*, *manifestar N*; se trata de nombres de <sentimiento> (Blanco, 2010). Otros aceptarán *N contraer*, *tratamiento para N*, *curarse de N*, *medicamento para N*... Se tratará de nombres de <enfermedad>, que podrán subdividirse en subclases observando la selección de predicados como *contagiar N a N*, *operar a N de N*, etc.

Entre los locativos, podremos destacar las <poblaciones>: *vivir en N*, *visitar N*, *llegar a N*, *habitante de N*. Entre los temporales, los nombres de <días de la semana> presentan una combinatoria específica: *el lunes por la mañana; a las 8.00 h del martes*, pero *#enero por la tarde*; *#a las 9.00 h del mes de marzo*.

El discente reparará, asimismo, en que algunas clases de locativos (*instituto*, *embajada*, *universidad*...) (*llegar a N*, *salir de N*, *encontrarse en N*...) admiten sistemáticamente una sintaxis propia de humanos de carácter colectivo (por ejemplo, pueden ser sujetos de verbos

15 Es importante remarcar, desde el punto de vista de la didáctica de las lenguas, que el *lex_umb* permite poner de relieve la importancia cuantitativa de algunas clases: algo más del 2 % de los N de *lex_umb* corresponden a la clase <prenda de vestir>, una proporción nada desdeñable. Otro 2 %, a <parte del cuerpo>, etc.

16 Obsérvese que empleamos *sustantivo*, *nombre* y la abreviación N de manera intercambiable.

de dicción o de pensamiento)[17]. Y algunas clases semánticas como los nombres de <texto> (*leer N*, *redactar N*, *traducir N*...) admiten, a un tiempo, combinatorias de diversos rasgos sintáctico-semánticos, como las correspondientes a concretos, abstractos y locativos: *romper una carta*, *tachar un párrafo*, *encontrar un artículo*, *saberse un poema*, *figurar en un libro*...

Es importante señalar que el docente no puede limitarse a observaciones dispersas, sino que debe disponer de una ontología semántica completa (Blanco, 2016), lo cual no implica en absoluto que deba enseñar dicha ontología, ni siquiera una parte significativa de la misma. Enseñar una lengua sin tener una ontología es como proponer la visita de una ciudad cuyo plano se desconoce. Enseñar una ontología, en cambio, sería como imponer la visita sistemática de cada calle de una ciudad. El buen guía tiene presente la totalidad del territorio por el que orientará a su grupo y es capaz de seleccionar las visitas oportunas según el momento adecuado, el tiempo disponible y los intereses de sus clientes.

El docente debería tener en su panoplia unas 400 clases semánticas, saber cuáles pueden ser más relevantes para los niveles iniciales y cuáles podrán irse seleccionando o introduciendo según el material que sus alumnos vayan a explotar. Ahora bien, este conocimiento no es, en absoluto, común. Existe, pues, un esfuerzo de formación de formadores que está, en gran medida, por hacer.

Precisemos que el docente no puede presentar simplemente una unidad léxica como *herida* (que, en nuestra clasificación, forma parte del nivel A2); en primer lugar, porque lo que debe enseñar no es una forma, sino un esquema proposicional: *herida de X por parte de Y(Humano o Animal) en Z(<parte del cuerpo> de X) con W(Inanimado Concreto)*.

Un esquema que retoma y refuerza rasgos y clases ya presentadas. También debe saber que *herida* pertenece a la clase <lesión>, que una nominalización de *cortar* (*corte*) es también una <lesión>, que un atenuativo de *herida* puede ser *arañazo* (ya para un nivel C2), etc.

La ontología debe hacerse descubrir, no presentarse como algo definitivo y acabado. De hecho, consideramos un posible error querer cerrar una ontología en un vocabulario orientado a la enseñanza de las lenguas.

El vocabulario ha de considerarse abierto e incompleto porque los taxones de una ontología presentan una importancia absolutamente disímil y, si bien es posible en gran medida

17 Presentar adecuadamente algunos casos de polisemia sistemática y de metonimias integradas (Kleiber, 1999: 104 y 121) puede ser muy útil para aumentar de manera importante la cobertura léxica de *lex_umb*, ya que este, en aras de la sencillez de manejo y de su aprovechamiento computacional en el marco de iRead4Skills, selecciona solo una unidad léxica (la más común) de cada vocablo.

prever un recorrido relativamente común para todos los discentes en A1 y A2, a partir de B1 los recorridos se van haciendo cada vez más variados. Basta ver cómo, en la práctica, se utilizan mucho menos los manuales de enseñanza de idiomas a partir de B1 que en los niveles inferiores, a no ser que el marco institucional esté muy bien definido (como, por ejemplo, en el marco de un examen de C1 orientado a la capacitación profesional).

3.3. Ámbitos de especialidad

Especificar el ámbito de especialidad para cada entrada de diccionario puede parecer redundante respecto a las clases semánticas o inaplicable para el léxico general. Pero no es así. En primer lugar, hay una diferencia entre formar parte o no de la terminología especializada (que es una propiedad diasistemática de una unidad léxica, cf. 4.2) y hacer referencia a un ámbito de la experiencia, que es una propiedad general a todas las unidades léxicas con significado no únicamente gramatical. Además, las ventajas de contar con esta información son importantes. En primer lugar, es una información que precisa las clases semánticas (no es totalmente independiente de ella, pero tampoco es completamente redundante). En segundo lugar, es operativa cuando se trata de clasificar un texto, ya que no resulta sencillo caracterizar un texto dado por las clases semánticas que se detectan en el mismo, pero, en cambio, no es difícil asignarlo a un ámbito de especialidad determinado a partir de la presencia de tres o cuatro unidades léxicas suficientemente específicas. Además, el conjunto clase semántica más ámbito de especialidad resulta robusto cuando se trata de desambiguar una unidad léxica (en especial un sustantivo, pero también ciertos verbos y adjetivos de carácter terminológico). Por añadidura, la codificación en ámbitos de especialidad requiere menos formación lingüística por parte del etiquetador que la codificación en clases semánticas.

Existen numerosas clasificaciones de ámbitos de especialidad; de manera empírica, pensamos que una lista de 600 ámbitos es una muy buena opción. Nuestra lista no está jerarquizada desde el punto de vista formal, pero sí desde el punto de vista conceptual, de tal manera que cada etiquetador llega al nivel de precisión que puede, sin perjuicio de que la etiqueta pueda ser revisada posteriormente. La relación de hiperonimia reposa enteramente en la forma de la etiqueta; *Derecho* es el hiperónimo de *Derecho administrativo*, *Derecho romano*, *Derecho civil*...

Para el etiquetado que hemos llevado a cabo en *lex_umb*, las indicaciones de especialidad resultan particularmente útiles, a fin de mejorar la caracterización de los nombres abstractos, ya que la limitación en el número de clases semánticas utilizadas hace que estas últimas no contengan información referencial destacada (Abst_Acción puede referirse a prácticamente todos los ámbitos). También son muy útiles para algunas clases de humanos (p. ej. <profesión>) o de locativos (p. ej. <establecimiento>): *hospital <establecimiento> Medicina*; *gimnasio <establecimiento> Deporte*; *universidad*

<*establecimiento*> *Enseñanza*. Nótese cómo la etiqueta de ámbito acompaña y precisa a la de clase semántica.

4. Fraseología y diasistemática

4.1. La fraseología

Las listas de vocabulario fundamental suelen ignorar casi por completo la fraseología[18]. Nuestro *lex_umb* tampoco contiene frasemas en su macroestructura. La disyuntiva es la siguiente: o se multiplica el tamaño del léxico (al menos por seis) o se encuentra un modo de dar cuenta parcialmente del fenómeno fraseológico que resulte operativo al menos en las etapas iniciales del aprendizaje.

Excepto en el caso de algunos frasemas gramaticales (*por qué*, *tan Adj/Adv que*) y de algunos pragmatemas (*buenos días*, *de nada*...), no es fácil llegar a un consenso sobre cuáles son los frasemas A1 y A2. Además, respecto a los lexemas, los frasemas parecen tener una diacronía más breve y son más sensibles a la variación diatópica (geográfica): *carnet de conducir* frente a *licencia de manejar*...

Proponemos optar, para el caso concreto que nos ocupa, por un tratamiento *ad hoc* de la fraseología que aproveche al máximo la competencia básica del lector y le ayude en los casos que puedan resultarle más problemáticos para la comprensión. Es importante distinguir cuidadosamente entre cuatro tipos de "expresiones fijas"[19].

4.1.1. Los frasemas pragmáticos

Los frasemas pragmáticos o pragmatemas (Blanco y Mejri, 2018) son de nivel superior a la unidad léxica; se trata, estrictamente hablando, de "frases hechas", a veces, incluso de pequeños textos. No tienen, pues, la entidad de lemas en un diccionario y deben tratarse en una lista aparte (*buenas tardes*, *feliz cumpleaños*, *consumir preferentemente antes de*...), porque exigen campos de descripción lexicográfica distintos a los que se aplican para las unidades léxicas; en particular, deben describirse la situación de comunicación y las coordenadas enunciativas (Blanco, 2014). Son fundamentales para la comunicación desde el nivel más básico del aprendizaje, pero predominan, en dicho nivel, los de carácter oral.

18 O bien presentan un número reducido de frasemas. Por ejemplo, el ELELex (CENTAL, 2014-2021), de poco más de 14 000 entradas, ofrece algo más de 600 frasemas. La proporción está más que invertida, ya que un locutor de una lengua dada conoce muchos más frasemas que lexemas.

19 Adaptamos la tipología propuesta en el marco de la lexicología explicativa y combinatoria (Mel' čuk, 2023).

4.1.2. Los frasemas léxicos

Los frasemas léxicos se dividen en frasemas completos (Mel'čuk, 2023: 64) y cuasi-frasemas (Mel'čuk, 2023: 67). Ambos son unidades léxicas y, en caso de tratarse, deben aparecer como entradas de macroestructura, pero ambos suelen, en un primer momento, dejarse algo de lado.

El significado de los frasemas completos no guarda relación con el de sus componentes (al menos, desde una perspectiva sincrónica): *buscarle tres pies al gato*, *echar de menos*, *talón de Aquiles*, *golpe de mano*, *estar por las nubes*, *poner de vuelta y media*, *de uvas a peras*, *mi media naranja*...

Cuando aparecen frasemas completos, estos deben tratarse como unidades léxicas a todos los efectos (con la particularidad de que son pluriverbales). Ahora bien, pese a ser importantes, no son ni tan numerosos ni tan frecuentes en discurso[20] como los otros tres tipos de frasemas. Tienen, además, la ventaja de que suelen suponer un bloqueo de la comprensión, con lo cual el discente no entiende, pero se suele dar cuenta de que no entiende (lo cual es importante). Pueden, pues, no introducirse de manera sistemática en un primer momento y elucidar los que vayan apareciendo en los materiales didácticos.

Existe una excepción importante: hay que presentar muy pronto una lista de algunos frasemas completos gramaticales de alta frecuencia: *sin embargo*, *a lo mejor*...

Los cuasifrasemas incluyen, en su significado global, el significado de, al menos, uno de sus componentes, o incluso de todos ellos, pero también incluyen significados suplementarios que no pueden deducirse de las unidades que los componen. El discente puede entenderlos al menos parcialmente: *llave inglesa*, *dar la mano*, *al aire libre*, *a mano (derecha, izquierda)*...

No hemos incluido, de momento, en *lex_umb* ni frasemas completos ni cuasi-frasemas. Es una limitación, pero es una limitación de la cual somos muy conscientes y que podremos paliar llegado el caso.

4.1.3. Los frasemas léxico-semánticos

Los frasemas léxico-semánticos o colocaciones tienen la característica de presentar dos partes muy diferenciadas. La primera (la base de la colocación) corresponde a una unidad léxica que no se ve condicionada por la fijación. La descripción lexicográfica del diccionario

20 Precisemos que son frecuentes en lengua (por consiguiente, muy numerosos si se desea repertoriarlos), pero su frecuencia de aparición en discurso es relativamente baja si lo comparamos a los otros tipos de frasemas.

le encaja, pues, sin ningún problema. Ahora bien, cuando a esa unidad léxica se le aplican ciertos tipos de predicados, el locutor nativo tiende a utilizar expresiones idiomáticas cuya selección léxica depende de la base. Así decimos: *un frío que pela*, *una memoria de elefante*, *una fiebre de caballo* (para expresar 'intensidad'), pero no *#un frío de elefante*, *#una memoria que pela*, etc. Existen decenas de miles de estas combinaciones, que van desde un fuerte carácter idiomático, como los ejemplos anteriores, hasta combinaciones que pueden parecer mucho más anodinas, pero que, de hecho, no son menos fraseológicas. Por ejemplo, para expresar 'bueno', tenemos *idea genial* o *peso ideal*, que son también colocaciones. Señalemos que el colocativo puede ser muy común en su uso léxico (*enfermedad grave*) o aparecer únicamente en el contexto de la colocación: *un error garrafal*, *creer a pie juntillas*.

Otro dato muy importante respecto a este tipo de frasemas es que la mayoría se centra en la expresión de unos pocos significados que son comunes a muchas (si no a todas) las lenguas; suele tratarse de predicados muy generales, como 'intenso', 'bueno', 'malo', 'empezar a', 'conjunto de'...). Otros expresan significados muy dependientes de la base (por ejemplo, 'tipo de'): *vino tinto*, *vino blanco* o *café solo*, *café cortado*, *café con leche*, *fruta del tiempo*, *año bisiesto*... Pero insistamos en que un *vino tinto* es un tipo de 'vino', mientras que *mi media naranja* no se refiere a una 'naranja', y *a mano derecha* no tiene directamente que ver con 'mano', etc.

Hay que aprovechar las características señaladas para tratar las colocaciones como información de microestructura[21]. De esta manera, se optimizan los recursos, se deja abierta la opción de seguir completando la cobertura y se entrena al discente a reconocer nuevos frasemas o, como mínimo, a barajar interpretaciones plausibles de nuevas colocaciones que pueda ir hallando.

Hemos abierto, hasta el momento, cinco campos de macroestructura: Intensivo, Atenuativo, Meliorativo, Peyorativo, Conjunto de y Verbo Soporte, que dan cuenta de secuencias como:

- 'intensivo': *abrir de par en par*, *conocer como la palma de su mano*, *comerse a besos*, *trabajar duro*, *amar apasionadamente*, *pena honda*, *gravemente enfermo*...
- 'atenuativo': *comer como un pajarito*, *cantidad irrisoria*...
- 'meliorativo': *respuesta correcta*, *marca premium*, *oportunidad de oro*...
- 'peyorativo': *vivir en la miseria*, *comida basura*, *negocio ruinoso*...

21 Es importante señalar que existe una fuerte correspondencia entre significados colocacionales y significados gramaticales (Blanco, 2006), y que, por tanto, se pueden realizar transferencias de aprendizaje; por ejemplo, entre los intensivos y los aumentativos morfológicos: *le dio una fuerte, violenta patada*; *le dio un patadón*.

— 'conjunto de': *flota de barcos*, *enjambre de abejas*, *banco de peces*...

— '∅'[22]: *dar un paseo*, *cometer un delito*, *correr un peligro, hacer una pregunta*...

Además, hemos reservado otro campo para las colocaciones TIPO DE (*vino joven*, *bistec poco hecho*, *coche deportivo*...), algo distintas, puesto que requieren ser precisadas semánticamente con una glosa.

En lo sucesivo, podrán irse abriendo nuevos campos para dar cuenta de otros significados colocacionales. Hagamos notar que algunos tipos de colocaciones se aplican solo a algunos tipos de base (p. ej., los verbos soporte se aplican solo a los nombres abstractos y a algunos adjetivos, 'conjunto de' se aplica a nombres, etc.). Además, la colocación suele poner de relieve uno de los significados del vocablo, ya que se refiere a una unidad léxica precisa. Para las colocaciones del español, existe un recurso excepcional, el diccionario REDES (Bosque, 2004).

4.2. La diasistemática

Algunas unidades léxicas están marcadas desde el punto de vista espacial, temporal o social. Algunas son arcaísmos; otras, neologismos; otras pertenecen a un registro familiar o culto; otras se usan solo, o predominantemente, en ciertas zonas geográficas del ámbito lingüístico. Para *lex_umb*, podemos dejar de lado la diacronía (o eje de variación temporal), ya que recogemos solo lexemas actualmente utilizados. No sucede así con la diatopía o eje espacial, puesto que la frecuencia de formas como *acá* variará mucho en función del área que consideremos. En la península ibérica, *acá* es una forma poco usada, pero es muy común en gran parte de la América hispana. En amplias zonas de América, la forma *carro* será frecuentísima, ya que designa al automóvil, pero en la península lo es mucho menos, ya que se utiliza *coche*; *carro* se reserva al cada vez más raro vehículo hipomóvil. Aunque *lex_umb* refleja sobre todo el habla peninsular, hemos considerado necesario destinar un campo específico a marcar como A1 las entradas que no lo son en Europa, pero sí en América; al fin y al cabo, estamos en un mundo globalizado en que el lector y espectador europeo consume productos americanos y viceversa[23]. Incluso en la conversación cotidiana el europeo se ve cada vez más confrontado a las variantes americanas también en suelo

22 Utilizamos el símbolo de 'conjunto vacío' porque los verbos soporte no tienen significado léxico, sino únicamente significados gramaticales.

23 Podemos tener americanismos por el referente (*dulce de leche*, *mate*...) o por el uso de la unidad léxica en sí (*pollera* vs. *falda*, *cancha* vs. *campo*). Particularmente en el primer caso, la globalización ayuda a que, en poco tiempo, la diatopía de nivel A1-B1 sea cada vez más accesible y es uno de los pocos aspectos del conocimiento del léxico en que nos ha parecido advertir un aumento de la competencia del discente medio en los últimos años.

peninsular, gracias a la cuantitativamente importante y cualitativamente valiosa inmigración que, procedente de toda Iberoamérica, se ha instalado en España.

Tenemos también el importante eje diastrático. Existen unidades léxicas de carácter familiar o hasta vulgar (*polla*, *tío*...) y unidades léxicas de carácter culto (*nefelibata*). En *lex_umb* se marcan solo las primeras, ya que las segundas se sitúan en niveles superiores de competencia. Finalmente, tenemos la variación diatécnica. Para esta última, en la medida en que indicamos el ámbito de especialidad, no hemos previsto un campo específico (sería informativo para marcar las unidades que pertenecen de manera estricta a una terminología, pero estas, si no se han generalizado, también son de niveles superiores a los recogidos en *lex_umb*).

Notemos que una unidad léxica puede estar marcada, a la vez, en distintos ejes diasistemáticos: *boludo* (familiar y americanismo rioplatense).

5. Conclusión

En este artículo hemos presentado brevemente los principios de elaboración de un léxico de nueva planta para los niveles A1, A2 y B1 del español, insistiendo en la importancia de concebir este diccionario como un recurso dinámico al servicio de la mejora de las competencias lingüísticas del discente, lo cual implica necesariamente una cierta formación metalingüística. Un recurso no significa gran cosa si no se sabe utilizar y, a pesar de la popularización de consignas como *aprender a aprender*, no parece que se haya hecho todavía suficiente hincapié en formar al discente (ni al docente[24]) en la utilización de recursos lingüísticos.

Nuestra metodología de elaboración de la macroestructura de un pequeño diccionario ha resultado muy costosa en términos de tiempo y constituye más un experimento lexicológico que un *modus operandi* para obtener vocabularios por niveles[25]. Respecto a las categorías

24 Si los programas de capacitación específica orientados a la enseñanza de las lenguas no garantizan una competencia metalingüística adecuada, esta se convertirá en una especie de saber arcano y no habrá posibilidad alguna de que se transmitan a los discentes ni siquiera las bases del análisis lingüístico. Y, sin embargo, estamos convencidos de que, incluso en estadios iniciales de la enseñanza, se pueden introducir ventajosamente nociones como las que hemos citado arriba, sin que ello suponga ningún *tour de force* inasumible. Es, por el contrario, una excelente manera de facilitar el aprendizaje, de dotar de autonomía al discente y de aprender a aprender. Reconozcamos, con pesar, que nuestra opinión es cada vez más minoritaria.

25 Aunque sus resultados no parecen en absoluto inadecuados. Estamos, en particular, gratamente sorprendidos por el hecho de que las particiones efectuadas han acabado correspondiendo de manera cuantitativamente muy aproximada a las cifras de vocabulario disponible estimadas para cada nivel (sin que nada en la metodología utilizada condicionase *a priori* las cifras obte-

de microestructura propuestas, pensamos que su difusión en el ámbito de la enseñanza de las lenguas resulta necesaria e importante.

Somos conscientes de que hemos introducido en unas pocas páginas un número tal vez algo excesivo de nociones. Nos parecía, sin embargo, necesario partir de este primer texto de carácter panorámico para, en lo sucesivo, tener una base en la que apoyar eventualmente otros artículos de carácter mucho más específicos que presenten en detalle al menos algunas de las nociones aquí abordadas. Queda, además, por abordar algún otro campo importante de microestructura, es el caso, en particular, de las derivaciones semánticas: con objeto de ampliar y estructurar el vocabulario del discente, es muy importante hacerle observar que *asesino* es el derivado semántico sujeto de *asesinar*, y *víctima*, su derivado objeto. Esto puede codificarse fácilmente en el diccionario. El parentesco morfológico ayudará en muchas relaciones de derivación semántica (*asesinar* > *asesino* o *vender* > *vendedor*), en otras, habrá solo relación semántica (*asesinar* > *víctima* o *vender* > *mercancía*/*producto* > *comprador*). En un primer momento, es prudente limitarse a los derivativos sujeto, objeto directo y objeto indirecto. Más adelante, pueden introducirse otros derivativos: instrumento (*tapar* > *tapa*/*tapón*; *sentarse* > *asiento*, *causativo* (*adelgazar* > *dieta*; *daño* > *siniestro*), etc.

No, la vaca de *lex_umb* sigue sin poder *mugir* y tampoco podemos *ordeñarla*. Habrá que esperar a niveles superiores para sentirse realmente cómodo en la granja de las palabras. Pero sabemos cómo y dónde esperar y ubicar esa información léxica (dos tipos de colocación) cuando llegue el momento y prever cuál será su semántica: 'producir N su sonido natural' y 'utilizar N de manera apropiada'.

6. Bibliografía citada

Blanco, Xavier, 2001: "Dictionnaires électroniques et traduction automatique espagnol-français", *Langages* 143, 49-69.

Blanco, Xavier, 2006: "Significacións gramaticais e sentidos colocacionais: ¿mais ca unha simple coincidencia?", *Cadernos de Fraseoloxía Galega* 8, 95-110.

nidas). La primera extracción efectuada en el diccionario de *flexsem* ha sido de 15 000 lemas, que corresponde bastante bien a la competencia pasiva de un nivel C2 (recordemos que los frasemas no están incluidos). La presencia de unas 750 entradas en A1, 1 500 en A2 y 3 000 en B1 también corresponde a las cantidades estimadas para estos niveles, aunque se podría pensar que deberían ser algo más elevadas en el caso del vocabulario pasivo. Nótese, no obstante, que este último está mucho más sujeto a variaciones interpersonales entre discentes que el vocabulario activo.

Blanco, Xavier, 2010: "Etiquetas semánticas de 'hecho' como género próximo en la definición lexicográfica", *Quaderns de filología. Estudis lingüístics* 15, 159-178.

Blanco, Xavier, 2014: "Inventaire lexicographique d'une sous-classe de phrasèmes délaissée: les pragmatèmes", *Cahiers de Lexicologie* 104, 133-153.

Blanco, Xavier, 2016: "A Hierarchy of Semantic Labels for Spanish Dictionaries" en Tatsiana Okrut, Yuras Hetsevich, Max Silberztein y Hanna Stanislavenka (eds.): *Automatic Processing of Natural-Language Electronic Texts with NooJ*, Cham: Springer, 66-73.

Blanco, Xavier, y Salah Mejri, 2018: *Les pragmatèmes*, Paris: Classiques Garnier.

Bosque, Ignacio (dir.), 2004: *REDES. Diccionario combinatorio del español*, Madrid: Ediciones SM.

Centre de Traitement Automatique du Langage [CENTAL], 2014-2021: *ELELex. A CEFR-graded lexical resource for Spanish as a foreign language* [https://cental.uclouvain.be/cefrlex/elelex].

Consejo de Europa, 2002: *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*, Madrid: Instituto Cervantes.

Garrigues, Mylène, 1992: "Dictionnaires hiérarchiques du français. Principes et méthode d'extraction", *Langue française* 96, 88-100.

Gougenheim, Georges, René Michéa, Paul Rivenc y Aurélien Sauvageot, 1956: *L'élaboration du français élémentaire: étude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris: Didier.

Gross, Gaston, 2013: *Manual de análisis lingüístico*, Barcelona: UOC.

Kleiber, Georges, 1999: *Problèmes de sémantique*, Villeneuve d'Ascq: Presses Universitaires du Septentrion.

Mel'čuk, Igor, 2023: *General Phraseology. Theorie and Practice*, Amsterdam / Philadelphia: John Benjamins.

Mel'čuk, Igor, y Alain Polguère, 2007: *Lexique actif du français*, Bruxelles: De Boeck & Larcier.

Polguère, Alain, 2016: *Lexicologie et sémantique lexicale*, Montréal: PUM.