Alonso Ramos, M. (2013): "Colocaciones, diccionario y corpus de aprendices". Casas Gómez, M. (dir.), Vela Sánchez, R. (ed.), Eugenio Coseriu, *in memoriam*, *XIV Jornadas de Lingüística*, Servicio de publicaciones Universidad de Cádiz, pp. 57-71.

Colocaciones, diccionario y corpus de aprendices

Margarita Alonso Ramos (Universidade da Coruña)

Resumen

En este trabajo establecemos las directrices de un entorno de aprendizaje en línea de colocaciones en español. Aunque, hoy en día la importancia de las colocaciones es generalmente reconocida, especialmente en la traducción y en el aprendizaje de una L2, consideramos que los recursos colocacionales son todavía escasos y poco refinados. Aquí nos centramos principalmente en dos tipos de recursos: en primer lugar, el *Diccionario de colocaciones del español* (DiCE) como parte de un entorno de aprendizaje y, en segundo lugar, en un corpus de aprendices (CEDEL2) como fuente de información que permita modelizar la producción colocacional de los aprendices de español.

Palabras clave: colocaciones, diccionarios, adquisición de L2, corpus de aprendices, ALAO

1. Introducción

El objeto de este trabajo son las colocaciones. Por *colocaciones* entendemos expresiones como las siguientes:

a. *dar PASEO / faire une* PROMENADE/ *fare una PASSEGIATA*
b. *FUMADOR empedernido / gros FUMEUR / FUMATORE incallito*
c. *diente de AJO* / *gousse d'AIL / spiccio d'AGLIO*

En estos ejemplos, tenemos una unidad léxica, aquí en versalita, que restringe la selección de la otra unidad léxica. Como explicaré un poco más tarde, una colocación es un tipo de unidad fraseológica, que consta de una base y un colocativo.

Cada día es más visible la necesidad de recursos léxicos que ayuden al aprendiz de lenguas en su quehacer multilingüístico. Como sabemos, aprender una palabra significa, también, aprender con qué otras palabras coocurre (Nation 2001). Aunque, hoy en día la importancia de las colocaciones es generalmente reconocida, especialmente en la traducción y en el aprendizaje de una L2, consideramos que los recursos colocacionales, en aquellos casos en que existen, están todavía muy lejos de ser refinados. En algunas lenguas, la falta de recursos es notable, comenzando por la ausencia de un diccionario de colocaciones. Esta es una de las razones que nos llevó a realizar la tarea de compilar una base de datos de colocaciones, fundamentada en la Lexicología explicativa y combinatoria (Mel'čuk et al. 1995): el *Diccionario de colocaciones del español* (DiCE), que presentaremos más tarde. Ahora bien, el desconocimiento que existe todavía sobre cuáles son las principales necesidades colocacionales de un aprendiz de español nos ha llevado a interesarnos por estudiar un corpus de aprendices y examinar con detenimiento qué colocaciones usan y de esas cuáles pueden considerarse correctas o incorrectas (Prieto et al. 2009, Alonso et al. 2010b, 2010c). El etiquetado de un corpus de aprendices nos ha servido para esbozar una primera tipología de errores colocacionales. Aquí nuestro objetivo es establecer las directrices para un entorno de aprendizaje en línea de las colocaciones en español como segunda lengua (ColocaTe). Pretendemos mostrar que es posible diseñar un entorno que sea más dinámico que los recursos existentes para otros idiomas. Mediante un sistema

dinámico, nos referimos a una interfaz integrada con acceso a una serie de recursos tales como diccionarios, corpus, incluida Internet, y una serie de actividades didácticas.

Nuestro trabajo se articula de la siguiente manera. Empezaremos por una breve presentación de la noción de colocación, tal como se entiende en nuestro marco teórico. A continuación, haremos algunas propuestas para un nuevo entorno que integre varios recursos y permita un fácil acceso al corpus. Después, presentaremos el DiCE, en su estado actual. Una vez presentado el diccionario, debemos ocuparnos del corpus de aprendices. Explicaremos la tipología de errores que hemos ideado para anotar el corpus de aprendices. Por último, formularemos las conclusiones y el plan de trabajo futuro.

2. Noción de colocación

Incluso si no existe una definición generalmente aceptada de colocación, en nuestro marco teórico se usa una definición operativa que se basa fundamentalmente en la coocurrencia léxica restringida. En la Lexicología Explicativa y Combinatoria (Mel'čuk et al. 1995), una colocación es un tipo de unidad fraseológica, que consta de dos unidades léxicas, la base y el colocativo. En las colocaciones se viola la propiedad de selección irrestricta, característica de los sintagmas libres: la base de la colocación se selecciona en función de su significado, mientras que el colocativo es seleccionado en función de la base. Digamos que en una colocación *AB*, la UL *A* es la base y la UL *B* es el colocativo. La base *A* es el núcleo semántico de la colocación y se selecciona libremente por su significado, mientras que el colocativo *B* es seleccionado en función de la base. La unidad léxica *B* puede significar 'B' normalmente o solo en combinación con *A*. En el primer caso, tenemos por ejemplo el nombre *loncha* que es el colocativo seleccionado por las bases *queso*, *jamón*, *chorizo* o *tocino* para expresar el significado 'porción de'; en el segundo caso, tenemos casos como *diente* que solo significa 'porción' cuando va con la base *ajo*. Ofrecemos a continuación algunos ejemplos más de colocaciones, en donde marcamos en versalitas la base de la colocación: del tipo nombre+adjetivo *ODIO mortal, LUCHA encarnizada*, *PACIENCIA infinita*, *CRIMEN atroz*, en donde el adjetivo expresa el sentido 'intenso'; algunas con el patrón verbo+nombre objeto como *despertar CURIOSIDAD*, *levantar SOSPECHAS*, *dar VERGÜENZA*, en donde el verbo expresa el sentido 'causar'.[1]

Una colocación no es una UL, sino que está formada por dos, la base y el colocativo. Puesto que el núcleo semántico es el de la base, las colocaciones deben ser descritas lexicográficamente en la entrada de la base. Por ejemplo, la colocación *diente de ajo* se describe en la entrada de la UL *ajo*. Una de las razones es el enfoque de síntesis o de producción que se defiende en nuestro marco teórico. A la hora de codificar, el hablante parte del sentido 'ajo' y busca la UL que exprese el sentido 'porción'. La mejor manera de describir esta información es en la entrada de la base, aunque por razones prácticas puede también reenviarse desde la entrada del colocativo. La decisión de si una unidad fraseológica dada es una colocación o una locución debe basarse en el análisis de su significado. Así, por ejemplo, si consideramos que *falda pantalón* significa 'falda con dos perneras', debemos describirlo como una colocación en la entrada lexicográfica de *falda*. En cambio, si consideramos que su definición se corresponde más con 'prenda que parece una falda pero tiene perneras', deberíamos tratarlo como una locución y asignarle una entrada lexicográfica[2].

[1] Para profundizar más sobre la noción de colocación, véase Alonso Ramos (2010).

[2] Estas dos definiciones de *falda pantalón* se corresponden, *grosso modo*, respectivamente con las definiciones aportadas por el DUE y por el DRAE. Si lo incluimos como una colocación es que lo estamos considerando como un tipo de falda. Según mi idiolecto, lo pluralizaría como *las faldas pantalón* y no ?*las faldas pantalones*.

3. Hacia un entorno de aprendizaje dinámico

En el ámbito de nuestro proyecto de investigación, estamos construyendo un entorno de aprendizaje de colocaciones, al que llamamos ColocaTe. Nuestras propuestas para ese entorno se pueden dividir en dos partes: la primera trata sobre el contenido del recurso y el segundo sobre su arquitectura.

3.1. Propuestas en cuanto al contenido

Ahora solo listaremos tres sugerencias que ejemplificaremos más adelante con el DiCE:

(1) La información en el diccionario debe ser introducida desde la base y los ejercicios debe centrarse en el colocativo.

(2) El diccionario debe contener descripciones semánticas y sintácticas de la colocación y los ejercicios tienen que proporcionar práctica en estos aspectos.

(3) Los ejemplos incluidos en un medio electrónico pueden y deben desempeñar un papel más importante. Es ampliamente conocido que los aprendices valoran los ejemplos más que cualquier otra cosa.

3.2. Propuestas en cuanto a la arquitectura

Proponemos una interfaz integrada que conecte varios recursos. Analizaremos a a continuación tres tipos de conexiones: 1) entre el módulo diccionario y el módulo didáctico; 2) entre corpus y diccionario; y 3) entre varios diccionarios.

3.2.1. Conexión entre el diccionario y las actividades didácticas

Con respecto a esta conexión, pensamos que el recurso debe:

1) Permitir al usuario completar el ejercicio consultando el diccionario. De hecho, pensamos que esta es la característica más atractiva de un entorno de aprendizaje de lenguas en donde estén varios recursos integrados.

2) Incluir las respuestas a los ejercicios. Es muy frustrante para un aprendiz no encontrar la respuesta en el diccionario a ejercicios planteados en el módulo de actividades didácticas (vid. Alonso Ramos 2006).

3) Poder ser usado de un modo diferente a cómo se usan los diccionarios y las actividades didácticas en papel. Por lo tanto, al hacer los ejercicios, en lugar de simplemente abrir la entrada de una palabra dada para buscar información, lo ideal es que el alumno pueda navegar por los datos utilizando un motor de búsqueda. Por ejemplo, si un ejercicio trata del verbo *guardar* como colocativo, la interfaz debe permitir al alumno consultar con qué bases este verbo puede formar una colocación.

4) Proporcionar no solo tests de verificación sino también contenidos. El vínculo entre los dos módulos, no sólo debe proporcionar ayuda en el momento de realización de los ejercicios, sino también en el proceso de corrección. Puesto que un programa de aprendizaje de lenguas asistido por el ordenador (ALAO) tiene la ventaja de proporcionar un *feedback* directo, sería más eficaz si en vez de dar simplemente "sí" o "no" como respuesta, ofreciera algunas referencias a la sección donde el diccionario trata con el problema en cuestión. El entorno, por lo tanto, debe incluir no solo pruebas de evaluación, sino también enseñanza o, dicho de otro modo, no solo debe servir para verificar si se sabe o no una colocación dada sino también para aprenderla.

5) Ofrecer un *feedback* flexible que permita respuestas alternativas, sobre todo en el área de colocaciones. Como sabemos, los criterios utilizados para determinar si una colocación es correcta o no, no siempre están bien definidos. Los juicios de aceptabilidad son muy sutiles entre combinaciones que un nativo puede decir con un

uso creativo, pero que quizás a un aprendiz no se le consienta. Pongamos el caso de una combinación como *admirador empedernido*. Un aprendiz de español podría consultar una herramienta como el DiCE para verificar si "existe" esta colocación. En este caso, no la encontrará, lo que no quiere decir que sea imposible. En el DiCE se proporcionan otros adjetivos para expresar la intensificación de *admirador* como *gran*, *rendido*, *devoto*, *confeso*, *ferviente, profundo* y quizás algún otro, que quizás parezcan más idiomáticos. Sin embargo, es cierto que el adjetivo *empedernido*, que estaba asociado a nombres que designan vicios o malos hábitos, está pasando a combinarse con otros nombres para expresar simplemente 'mucho' o 'muy intenso'.

6) Registrar los resultados de los ejercicios hechos por los aprendices. Esta idea ya ha sido puesta en práctica con excelentes resultados en Alfalex, la plataforma de ejercicios vinculada a la *Base lexicale du français* (Verlinde et al. 2003, 2010). El usuario puede buscar los errores que ha hecho y puede pedir al sistema que le facilite más actividades que se dirijan al mismo problema.

3.2.2. Conexión entre diccionario y corpus

Según Kilgarriff (2005), hay dos maneras de modelar este vínculo: 1) "poner el corpus en el diccionario" (PCED), estrategia seguida por la actual Lexicografía basada en corpus; 2) "poner el diccionario en el corpus" (PDEC), postura que sería la defendida por los que codifican la información en el corpus, tal y como se hace en algunos trabajos de desambiguación semántica automática o en cualquier tarea de anotación de corpus que necesite información léxica. Sin embargo, también es posible adoptar otra manera de hacer interactuar corpus y diccionario. Dado que el diccionario es una base de datos en la que extractos del corpus explotado pasan a ser registrados en alguno de los campos de la base, puede decirse que el diccionario contiene también un corpus; un corpus que puede ser separado del resto de la información incluida en el diccionario. Por lo tanto, se puede acceder al corpus a través del diccionario, en lugar de utilizar un corpus anotado externo.

Con fines didácticos, se debe agregar una herramienta de búsqueda. Los usuarios utilizarán esta herramienta cuando no están necesariamente interesados en consultar la entrada de una base. Por ejemplo, si un aprendiz desea saber si el nombre *admiración* tiene un determinante o no cuando se va con el verbo *tener*, en vez de ir a la entrada de *admiración* y desplazarse buscando la información requerida, sería más rápido y más útil poner en marcha una herramienta de búsqueda que navegue por todo el corpus incluido en el diccionario. La herramienta busca la coocurrencia entre *admiración* y *tener* y puesto que los ejemplos de *tener* se agrupan por el lema, el corpus no requiere ser etiquetado.

También defendemos los beneficios de vincular la interfaz con un motor de búsqueda como Google y con un corpus de referencia como el CREA. Además de proporcionar a los aprendices con una mayor autonomía, un corpus externo les permite verificar si el hecho de que una colocación no esté incluida en el diccionario es simplemente una omisión o si esa presunta combinación efectivamente no existe, como mencionamos arriba a propósito de *admirador empedernido*. Puesto que la lengua está en continua evolución, la no inclusión de una colocación en el diccionario no debe implicar que esa combinación sea incorrecta. La interfaz puede añadir un cuadro de diálogo de búsqueda en Google advirtiendo a los aprendices que puedan encontrar ejemplos no estándar que no son aceptables en la lengua escrita (Milton 2006). En este caso concreto, la consulta desde Google de la combinación "admirador empedernido", le daría 1510 ocurrencias, lo que no es muy alto (enero de 2012). Desde el CREA, la consulta "admirador dist/5 empedernido" devuelve un solo ejemplo. Sería interesante la

posibilidad de incluir desde el DiCE la interpretación de estas consultas a otros corpus, con el fin de facilitar la tarea al usuario.

3.2.3. Conexión con otros diccionarios

Aunque la interfaz se centra en las colocaciones, sería más útil si el usuario pudiera conectarse con otros diccionarios monolingües y bilingües. La edición de 2009 del *Oxford Collocations Dictionary* es la prueba de que la combinación de un diccionario de colocaciones y un diccionario monolingüe lo convierten en un mejor producto. Pensamos que un entorno de aprendizaje útil tiene que parecerse a una estación de trabajo similares a las que utilizan los traductores.

Por ejemplo, una de las actividades didácticas propuestas por Higueras (2006) se centra en el hecho de que un colocativo puede o no tener el mismo significado en una colocación dada que en una combinación libre. Así, por ejemplo, el sustantivo *duda* elige el verbo colocativo *aclarar*. Este verbo tiene diferentes significados en diferentes combinaciones, por ejemplo, *aclarar la ropa* y *aclarar la voz*. Para poder hacer el ejercicio correctamente, los aprendices tienen que consultar varios diccionarios. Si son suficientemente competentes en la lengua, un diccionario monolingüe en la lengua meta puede ser suficiente, pero a veces los diccionarios monolingües, especialmente en español, no proporcionan ninguna información sobre el significado de las unidades léxicas en la colocación. De hecho, el sentido de *aclarar* en colocación con *duda* no se describe en los diccionarios monolingües en español. Los aprendices que consultan un diccionario bilingüe español-inglés encontrarán el verbo *to clarify* lo que puede resolver algunos aspectos de este ejercicio, pero el verbo *to clarify* no es el colocativo equivalente para el nombre *doubt*: en inglés, este nombre se combina mejor con *to clear up*, *to dispel*, *to remove* or *to resolve*. Por lo tanto, un diccionario de colocaciones de la lengua del alumno podría ser un recurso adicional útil.

A continuación, mostraremos cómo algunas de estas propuestas han sido puestas en práctica en un recurso colocacional para el español, el DiCE.

4 DiCE: presentación del estado actual

A diferencia de otros recursos léxicos que son una réplica electrónica de los diccionarios en papel, el DiCE ha sido concebido como una base de datos léxicos electrónica, disponible en la web desde 2004 con constantes modificaciones y mejoras[3]. La arquitectura de un diccionario electrónico es necesariamente una red, no una lista. En su estado actual, el DiCE contiene alrededor de veinte mil relaciones léxicas. Empezaré con una breve presentación de la arquitectura del DiCE. Se compone de tres componentes principales: el diccionario propiamente dicho, las actividades didácticas (versión beta) y el componente de búsquedas avanzadas

Empecemos por el diccionario propiamente dicho. El DiCE se caracteriza por las siguientes propiedades: (1) cada colocación recibe una descripción semántica y sintáctica; (2) se ilustra cada colocación con varios ejemplos, la mayoría de los cuales provienen del CREA; (3) se organiza por campos semánticos, y (4) su orientación hacia la producción.

La estructura de los datos se organiza en torno a los lemas. Cada lema agrupa varias unidades léxicas (UL) diferentes. Para cada UL, el usuario puede consultar la información semántica o combinatoria correspondiente. Con respecto a la información semántica**,** para cada UL, la entrada dispone lo siguiente: a) la *etiqueta semántica* que

[3] El DiCE se puede consultar en la dirección:www.dicesp.com. Para alguna presentación más detallada, vid. Alonso Ramos (2008), (2010), Alonso et al. (2010a) y Vincze et al (2011b).

representa el significado genérico, b) la *estructura actancial* que representa a los participantes de la situación designada por el sustantivo; c) ejemplos de corpus, con mayor frecuencia del CREA, y d) cuasisinónimos y cuasi-antónimos. Con respecto a la información combinatoria**,** nos centraremos aquí exclusivamente en la información léxica combinatoria, es decir, en las colocaciones. En todos los diccionarios procedentes de la Lexicología Explicativa y Combinatoria, las colocaciones están descritas semántica y sintácticamente en términos de *funciones léxicas.* Una FL codifica la relación entre dos unidades léxicas en donde una de ellas (la base de la colocación) controla la elección léxica de la otra (el colocativo). Por ejemplo, la FL `Magn` codifica la relación entre los siguientes pares de nombre-adjetivo: *honda pena terrible vergüenza,* y *ferviente admiración.* Los tres adjetivos (los colocativos) son seleccionados para expresar, en combinación con el sustantivo correspondiente, el mismo significado 'intenso'. Aunque las FFLL sean los cimientos de toda la red colocacional de los diccionarios procedentes de este marco teórico, tanto en el DiCE como en LAF (Mel'čuk y Polguère 2007), hemos optado por dar más visibilidad a la glosa semántica que a la FL.

En cuanto a la ejemplificación**,** todas las colocaciones son apoyadas por ejemplos procedentes de corpus. Por lo tanto, aunque las FFLL constituyen una red que sirve de guía para la búsqueda de las colocaciones en el corpus, el procedimiento se basa en los datos. Esto permite al DiCE ser utilizado como un corpus de colocaciones, pero un corpus que se describe y se explica. Podríamos decir que el corpus en el DiCE, es un corpus sin anotar, pero enriquecido puesto que los ejemplos de cada colocación están asociados a una FL con una base y un colocativo, ambos lematizados. Por lo tanto, el corpus "crudo" (*raw*), seleccionado por el lexicógrafo, se enriquece en cuanto pasa a formar parte del DiCE. Así, por ejemplo, la colocación *dar opinión* del extracto del corpus *Anteriormente ya* di mi opinión sobre *este producto*, al ir asociada con la información de la FL, sabemos: 1) que el verbo significa 'expresar', por lo tanto, estamos desambiguando el polisémico verbo *dar*; 2) que el sujeto de ese verbo es el primer argumento del nombre *opinión*, el "Cognizer" en los términos de FrameNet; 3) que el complemento preposicional es el segundo argumento del nombre, el "Topic".

El corpus de colocaciones contenido en el DiCE, como recurso independiente, puede ser anotado y analizado sintácticamente para ser explotado como un *treebank* de colocaciones, que sirva a la identificación automática de colocaciones. Los primeros pasos a este respecto ya han sido dados en el proyecto de investigación presente.

Examinemos ahora someramente el módulo didáctico**.** Se encuentra todavía en una etapa preliminar, pero aún así, la semántica y la descripción sintáctica de las colocaciones en el DiCE permite ejercicios más interesantes que los que se basan exclusivamente en emparejar bases y colocativos. Por ejemplo, podemos dar a escoger entre diferentes adjetivos que se corresponde con un sentido dado. Si el usuario elige una respuesta incorrecta, le reenvía al DiCE en donde puede encontrar la descripción de la colocación *alegría pasajera*. Véase Fig. 1.

Ahora, pasemos al tercer componente**.** Las *consultas avanzadas* sirven principalmente para hacer búsquedas específicas, en lugar de hacer consultas de todas las colocaciones de una unidad léxica. Con respecto a la conexión corpus-diccionario, la estructura del DiCE permite la inclusión de motor de búsqueda que navega por todo el corpus de colocaciones. A diferencia de una consulta en la web, el corpus del diccionario ya ha sido tratado y desambiguado, como hemos visto. Por lo tanto, cuando los aprendices quieren saber si el nombre *admiración* se combina con el *verbo* tener, la herramienta de búsqueda devuelve los ejemplos separados en dos grupos, clasificados de acuerdo con las diferentes FFLL: *tener admiración por alguien* 'sentir admiración',

codificada por la FL $Oper_1$ y *tener la admiración de alguien* 'ser objeto de la admiración de alguien', codificada por el LF $Oper_2$. Véase Fig.2.

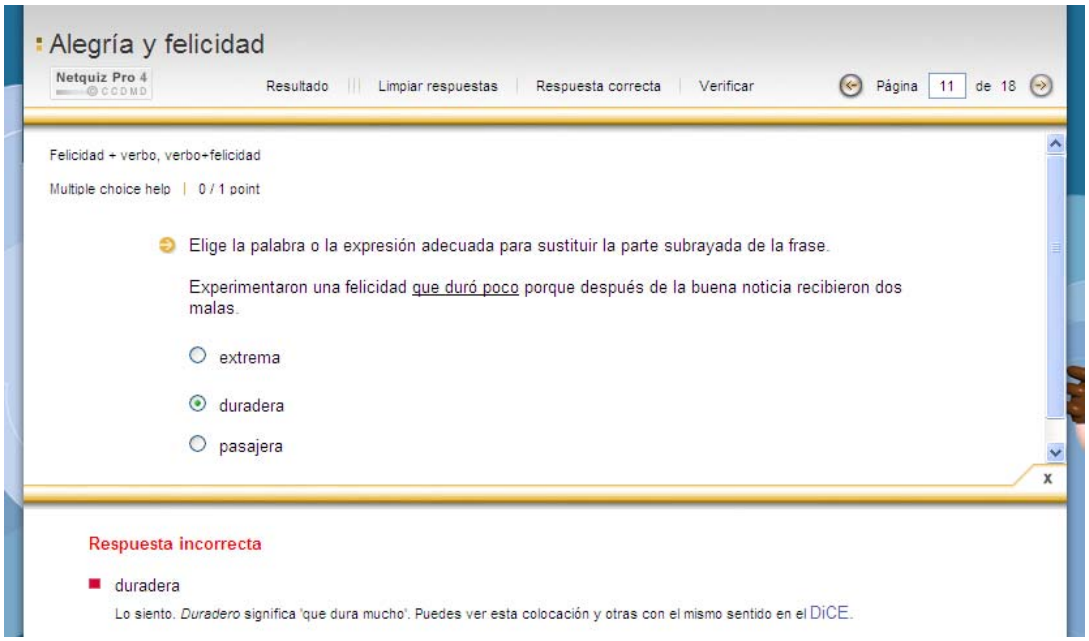

Fig.1 Actividades didácticas del DiCE

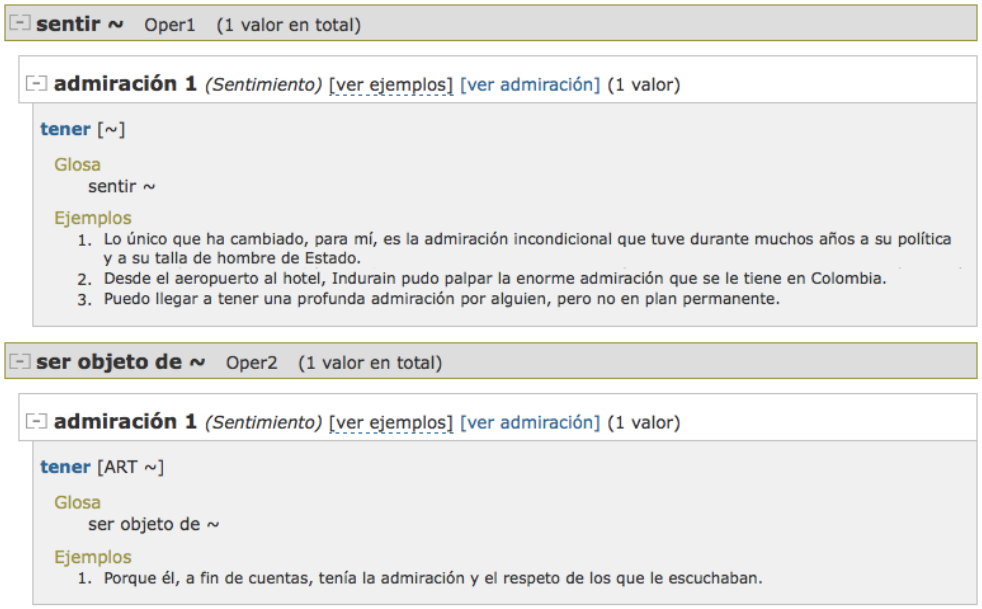

Fig.2 Resultado de una consulta avanzada

Hasta ahora hemos hablado de las colocaciones en un diccionario. Pasemos ahora a examinar las colocaciones en un corpus, más en particular, un corpus de aprendices.

5. Colocaciones en un corpus de aprendices

Para poder ofrecer buenas herramientas de aprendizaje, debemos saber qué tipos de errores colocacionales cometen los aprendices. Por esta razón, nos interesamos por

CEDEL2 (Lozano 2009). Ese corpus ha sido elaborado por un grupo de la Universidad Autónoma de Madrid[4]. Contiene redacciones en español escritas por hablantes nativos de inglés sobre un conjunto predefinido de temas. Los textos están clasificados con respecto al nivel de conocimiento del español de los autores. La muestra en la que se basa nuestro estudio contiene textos de aprendices con nivel de español intermedio o avanzado.

Los errores colocacionales son muy variados. Por ejemplo, en s*alvar dinero* (por *ahorrar dinero*) el aprendiz toma una unidad léxica existente en español, para el colocativo, pero con otro significado; en *recibir un llamo,* el error está en la base: *llamo* no existe; en *asistir la universidad*, vemos un error gramatical, donde el alumno trata el verbo *asistir* como un verbo transitivo. Por lo tanto, se necesita una clasificación más detallada de errores colocacionales para ofrecer a los alumnos material de aprendizaje más específico (y por lo tanto más eficaz), y para facilitar el desarrollo de técnicas para la corrección automática de errores colocacionales en los escritos de los aprendices.

En el ámbito del proyecto ColocaTe, desarrollamos una tipología de errores colocacionales que distingue tres dimensiones paralelas. La primera dimensión (la *localización*) nos dice si el error se refiere a la colocación en su conjunto o en uno de los miembros (la base o el colocativo); la segunda dimensión modela los errores desde un punto de vista descriptivo; y la tercera ofrece el análisis explicativo[5].

Para llevar a cabo la anotación del corpus con esta tipología, adaptamos *Knowtator,* como herramienta de anotación. En lo que sigue, presentaremos la tipología de errores colocacionales, las ideas principales del procedimiento de anotación con Knowtator y algunos resultados preliminares derivados de nuestra anotación.

A nivel descriptivo**,** una colocación puede ser errónea desde una perspectiva léxica, gramatical o de registro. Ahora sólo quiero hacer hincapié en que esta dimensión tiene en cuenta los diferentes tipos de información que un aprendiz tiene que saber para poder utilizar una colocación correctamente: tiene que saber cuál es la buena unidad léxica (base o colocativo), pero también tiene que saber cómo usarlas con corrección gramatical y con adecuación contextual. Hemos decidido explicitar las posibles causas de error, especialmente en el caso de errores léxicos. Creemos que una dimensión explicativa de la tipología de errores es de gran utilidad para diseñar material didáctico. Para permitir una mayor flexibilidad, prevemos la posibilidad de que más de una causa se asigna a un solo error. La distinción más genérica es entre error *interlingual* (o transferencia L1-L2) y el error *intralingual* L2. Esta distinción se refiere a los tres principales tipos de errores que se introducen en el nivel descriptivo: léxico, gramatical y de registro. En cuanto a los errores de gramática y de registro, no hacemos ninguna distinción explicativa adicional. Por ejemplo, un error gramatical de régimen puede ser descrito como interlingual o como intralingual. En el primer caso, podemos observar la influencia del inglés; por ejemplo, en *terminé escuela,* lit.'I finished school'. En el segundo caso, el régimen erróneo no se puede atribuir directamente a la L1, por ejemplo, *montar el autobús.* En este caso, consideramos que es intralingual.

Veamos algunos ejemplos de errores léxicos interlinguales**.** Los errores interlinguales se dividen en dos subclases: (a) *la importación:* el aprendiz crea en L2 una unidad léxica a partir de otra unidad léxica en su lengua materna, lo más a menudo adaptándola a la forma de L2 (como en *recibir un llamo*), (b) *la extensión:* el aprendiz extiende el significado de una LU existente en la L2. En este caso, tenemos una

[4] El corpus de aprendices bajo estudio es un fragmento de CEDEL2 Para más información, véase la página web: http://www.uam.es/proyectosinv/woslac/cedel2.htm

[5] La tipología de errores así como el proceso de anotación del corpus han sido descritos con mayor detalle en Alonso Ramos et al. (2010b), (2010c) y Vincze et al. (2011a).

sustitución: cambio incorrecto de un colocativo o una base por otra palabra que existe en español. En muchos casos de extensión, la UL en L1 es una traducción válida de la LU en L2, pero con un significado diferente al pretendido. Por ejemplo, *gastar el año* en lugar de *pasar el año,* donde *gastar* es elegido por una posible traducción de *spend.* Un error de extensión se produce a menudo a causa de que se utiliza una UL en L2 por su similitud fonética con la forma equivalente en L1: p. ej *maternal* en *lengua maternal* en lugar de *materna*. También cuando el uso de una UL en L2 se evita precisamente porque parece formalmente muy similar a su equivalente en L1 - lo que puede considerarse un caso de hipercorrección: *atender* en *atender el teléfono* es descartado por el aprendiz en favor de *acudir*: *acudir el teléfono,* ya que *atender* parece muy similar al inglés *to attend.*

En cuanto a los errores léxicos intralinguales, distinguimos tres subclases: (a) *la derivación errónea:* el aprendiz produce una forma inexistente en la L2 como resultado de un proceso de derivación errónea, por analogía con otro tipo de L2 (cf. *enseñanza se**g**undaria,* en vez de *secundaria),* (b) *la generalización:* el aprendiz elige una UL más vaga o más genérica de lo necesario (cf. *hacer citas,* en lugar de *concertar citas*). Este es un ejemplo de un error con dos interpretaciones posibles, ya que no está claro si el aprendiz utiliza el verbo *hacer,* debido a la influencia de la L1 (inglés, en este caso) (como en *I'd like to make an appointment to see the doctor, please*) o simplemente, porque *hacer* a menudo funciona como un verbo de apoyo en español; (c) *la elección errónea:* el aprendiz selecciona una UL errónea, sin una razón clara y sin la intervención de la L1 *(escribir el examen*, en lugar *de hacer el examen*).

Los errores gramaticales más frecuentes conciernen el régimen - como, por ejemplo, el uso como transitivo de un verbo que requiere una preposición: *asisto la universidad,* en lugar de *asisto a la universidad*.

En la siguiente figura, se muestra una captura de pantalla con la herramienta de anotación. Knowtator nos permite definir un esquema de anotación que se adapta a la tipología de errores y la tipología de colocaciones (correctas) dada por las funciones léxicas. Además de la anotación de los errores, el anotador también puede anotar colocaciones correctas que se encuentran en el corpus con la información correspondiente de la Función léxica. La figura muestra un fragmento del corpus en el que se etiquetan colocaciones correctas (en verde) e incorrectas (en rojo). El zoom de la figura se centra en el proceso de etiquetado de la colocación errónea *comanda el respeto,* en lugar de *imponer respeto* con la etiqueta “léxico - extensión por semejanza fonética", de la dimensión explicativa y "sustitución" como etiqueta de la dimensión descriptiva.

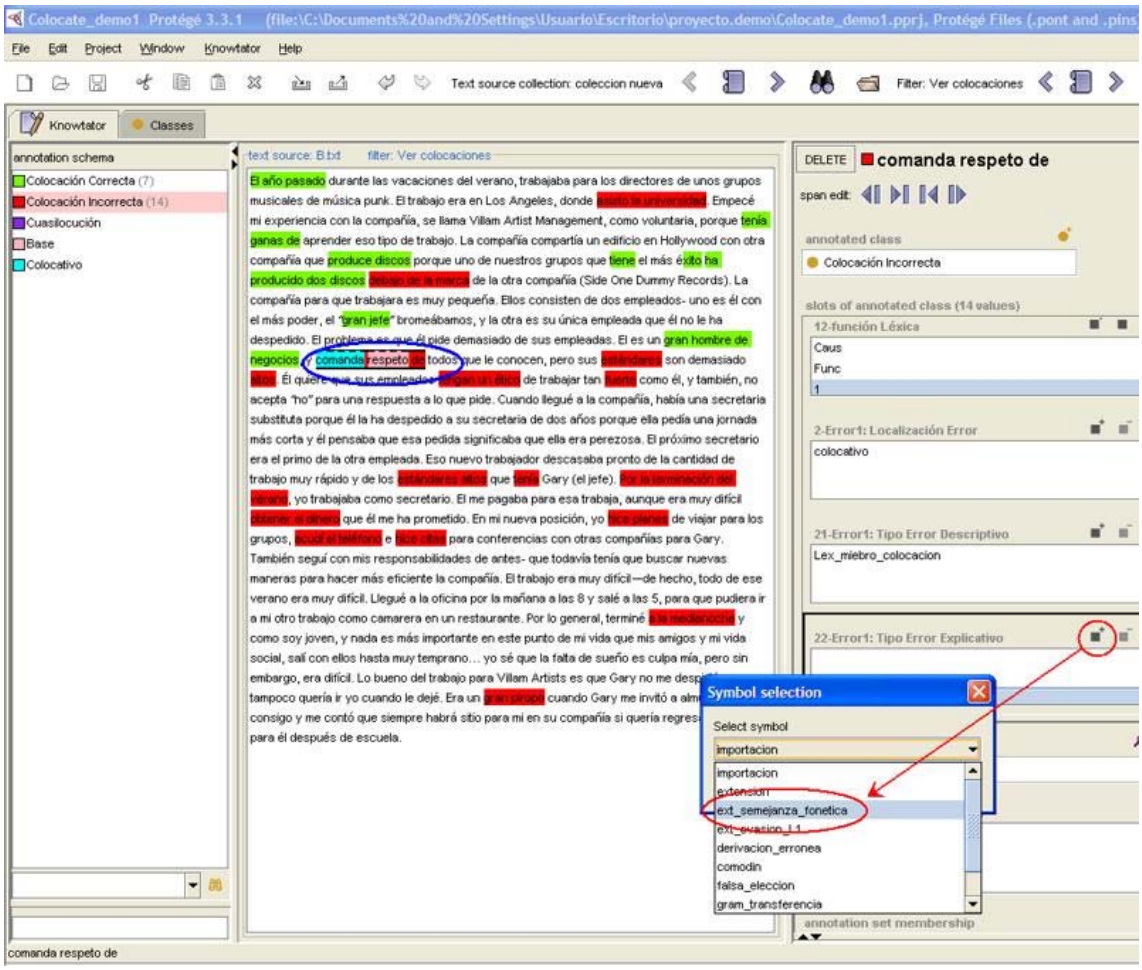

Fig.3 El proceso de anotación con Knowtator

En este momento contamos con un corpus anotado compuesto por 100 redacciones de nivel intermedio y avanzado, con un total de 46.000 palabras. Del total de colocaciones etiquetadas (1864), un 76% son correctas y el 24% son incorrectas. Aunque la mayoría de errores son léxicos, encontramos un 44% de colocaciones afectadas por errores gramaticales. El 61% de errores afectan al colocativo, el 20,5% a la base y el 18,5% a la colocación entera. El error más frecuente de tipo gramatical concierne al régimen, un 40%. En otras palabras, los aprendices hacen diferentes tipos de errores colocacionales. Estos errores son demasiado diferentes para ser simplemente etiquetados como "errores léxicos".

6 Conclusiones

Para concluir, me gustaría resumir brevemente los puntos principales en dos grandes bloques. Comenzando por el final, vayamos primero al punto, las colocaciones en un corpus de aprendices. Hemos tratado de mostrar que los errores colocacionales están lejos de ser homogéneos. Por el contrario, una tipología fina de errores colocacionales como la ofrecida aquí puede ser útil para proporcionar material didáctico que permita un aprendizaje de la lenguaje activo. Teniendo en cuenta que nuestra tipología es independiente de la lengua, esperamos que sea de utilidad para la comunidad de ALAO en general.

El otro bloque, colocaciones en un diccionario, hemos querido señalar que hay otra manera de concebir los recursos léxicos. Somos consciente de que un entorno dinámico de aprendizaje para las colocaciones es un objetivo de investigación a medio plazo, pero creemos que el DiCE está en el buen camino. Nuestro marco teórico facilita muchas de las tareas. Puesto que las FFLL constituyen un lenguaje formal, sistematizan la información colocacional, lo que facilita la integración del diccionario y los módulos didácticos en un entorno de lengua de aprendizaje en línea.

Con todo, queda trabajo por hacer**.** Además de seguir con el DiCE, el módulo didáctico debe ser desarrollado por poner las diferentes propuestas discutidas anteriormente en la práctica. Por ejemplo, sería interesante generar y corregir los ejercicios automáticamente, a partir de la información contenida en el DiCE. Y, por supuesto, continuar con la anotación del corpus de aprendices luchando por lograr el máximo acuerdo entre los anotadores.

Agradecimientos

Este trabajo ha sido realizado en el marco de los proyectos de investigación, parcialmente subvencionados por fondos FEDER y por el Gobierno de España: FFI2008-06479-C02-01 y FFI2011-30219-C02-01. Aprovecho también la ocasión para agradecer a los organizadores de las Jornadas que tuvieron lugar en 2011 en la Universidad de Cádiz.

Referencias bibliográficas

ALONSO RAMOS, M. (2008): "Papel de los diccionarios de colocaciones en la enseñanza de español como L2", *Proceedings of the XIII EURALEX International Congress*, Barcelona, 1215-1230.

ALONSO RAMOS, M. (2009): "Hacia un nuevo recurso léxico: ¿fusión entre corpus y diccionario", Cantos Gómez, P., Sánchez Pérez, A. (eds): *A Survey of Corpus-based Research. Panorama de investigaciones basadas en corpus*. Murcia: AELINCO; 1191-1207.

ALONSO RAMOS, M. (2010): "No importa si la llamas o no colocación, descríbela". Mellado, C. et al. (eds.), *La fraseografía del S. XXI: Nuevas propuestas para el español y el alemán*, 55-80. Berlin: Frank & Timme.

ALONSO, M; NISHIKAWA, A; VINCZE, O. (2010a): "DiCE in the web: An online Spanish collocation dictionary", S. Granger, M. Paquot (eds.), *eLexicography in the 21st century: New Challenges, New Applications. Proceedings of eLex 2009*, Cahiers du Cental 7, Louvain-la-Neuve, Presses universitaires de Louvain, pp. 367-368.

ALONSO, M.; WANNER, L.; VÁZQUEZ, N.; VINCZE, O.; MOSQUEIRA, E.; PRIETO, S. (2010b): "Tagging collocations for learners", S. Granger, M. Paquot (eds.), *eLexicograpy in the 21st century: New Challenges, New Applications*.

Proceedings of eLex 2009, Cahiers du Cental 7, Louvain-la-Neuve, Presses universitaires de Louvain, pp. 369-374.

ALONSO RAMOS, M., L. WANNER, O. VINCZE, G. CASAMAYOR, N. VÁZQUEZ, E. MOSQUEIRA, S. PRIETO (2010c): "Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora", *7th International Conference on Language Resources and Evaluation (LREC)*, La Valetta, Malta, pp. 3209-3214.

CROWTHER, J., DIGNEN, S., LEA, D. (eds.) (2009): *Oxford Collocations Dictionary for Students of English.* Oxford: Oxford University Press.

HIGUERAS, M. (2006): *Las colocaciones y su enseñanza en la clase de ELE*. Madrid: Arco/libros

KILGARRIFF, A. (2005): "Putting the Corpus into the Dictionary", *Proceedings MEANING Workshop*, Trento.

LOZANO, C**.** (2009): "*CEDEL2:* Corpus Escrito del Español L2", C. M. Bretones Callejas et al. (eds.) *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente.* Almería: Universidad de Almería, pp. 80-93.

MEL'ČUK, I., A. CLAS, A. POLGUERE (1995) : *Introduction à la lexicologie explicative et combinatoire,* Louvain-la-Neuve : Duculot.

MEL'ČUK, I., POLGUÈRE, A. (2007): *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20.000 dérivations sémantiques et collocations du français*. Louvain-la-Neuve: de Boeck Duculot.

MILTON, J. (2006): "Resource-Rich Web-Based Feedback: Helping Learners Become Independent Writers", K. Hyland, F. Hyland (eds.) *Feedback in Second Language Writing: Contexts and Issues*, Cambridge, Cambridge University Press.

NATION, I.S.P. (2001): *Learning Vocabulary in Another Language*, Cambridge: Cambridge University Press

PRIETO, S., MOSQUEIRA, E, VÁZQUEZ, N. (2009): "Córpora y enseñanza de lenguas: se buscan colocaciones", Cantos Gómez, P., Sánchez Pérez, A. (eds.): *A Survey of Corpus-based Research. Panorama de investigaciones basadas en corpus*. Murcia: AELINCO; 366-373.

VERLINDE, S., SELVA, T., BINON, J. (2003) : "Alfalex : un environment d'apprentissage du vocabulaire français en ligne, interactif et automatisé ", *Romaneske* 28, 1, 42-62.

VERLINDE, S., PAULUSSEN, H., SLOOTMAEKERS, A., DE WACHTER, L. (2010): " La conception de didacticiels intégrés d'aide à la lecture, à la traduction et à la rédaction ". *Revue Française de Linguistique Appliquée,* 15 (2): 53-65.

VINCZE, O., M. ALONSO RAMOS, E. MOSQUEIRA SUÁREZ, S. PRIETO GONZÁLEZ. (2011a): "Exploiting a learner corpus for the development of a CALL environment for learning Spanish collocations", Kosem, I. y K. Kosem, eds. *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011*. Ljubljana: Trojina, Institute for Applied Slovene Studies, 280-285.

VINCZE, O., MOSQUEIRA. E., ALONSO RAMOS. M. (2011b): "An online collocation dictionary of Spanish", Boguslavsky, I. y L. Wanner, eds. *Proceedings of the 5th International Conference on Meaning-Text Theory Barcelona, September 8-9, 2011*, 275-286.