HACIA UN NUEVO RECURSO LÉXICO: ¿FUSIÓN ENTRE CORPUS Y DICCIONARIO?

Margarita Ramos
Universidade da Coruña

1. Introducción

En los últimos años el concepto de diccionario está cambiando en su relación con el corpus y, como mostraremos, hay indicios de que el diccionario y el corpus se empiezan a (con)fundir. La interacción entre corpus y diccionario es continua. Por un lado, es impensable actualmente abordar una empresa lexicográfica sin apoyarse en un corpus. El corpus no solo sirve para proporcionar ejemplos a la información ya incluida en el diccionario, sino que la práctica actual en Lexicografía es que el corpus conduce la búsqueda de información que se incluye en el diccionario. Como veremos más tarde, actualmente la Lexicografía no solo está basada en el corpus sino conducida por él (Krishnamurty 2008). Por otro lado, la anotación de corpus ocupa un lugar cada vez más prominente en los estudios lingüísticos y para anotar un corpus, bien sea morfológica, sintáctica o semánticamente, se necesita de una manera o de otra un diccionario. La influencia entre ambas herramientas se ve facilitada, además, por los medios electrónicos puesto que los diccionarios modernos se conciben como bases de datos en soporte informático. Por esa razón, el diccionario se asimila cada vez más a una plataforma de acceso a la información léxica y esa información puede estar en (o ser) el propio corpus. En esta comunicación, nos proponemos analizar cómo está evolucionando la relación entre corpus y diccionario de tal modo que de interacción se empieza a pasar a solapamiento. La tesis que vamos a defender es que el diccionario debe contener un corpus pero el segundo no puede sustituir al primero. Por tanto, defenderemos el papel importante del corpus pero sometido al diccionario.

Para abordar la interacción entre corpus y diccionario, puede resultar útil la distinción establecida por Kilgarriff (2005) entre las dos vías siguientes: 1) "poner el corpus en el diccionario" (PCED), que representa la actual Lexicografía basada en corpus; 2) "poner el diccionario en el corpus" (PDEC), 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. En la sección siguiente, exploraremos con algún detenimiento estas dos vías para llegar a la sección 3 en donde abordaremos más en detalle este proceso de hibridación entre corpus y diccionario que ya estamos anunciando. Con el fin de ejemplificar ese emergente concepto de diccionario como una interfaz de acceso a la información contenida en el corpus, nos detendremos en mostrar un caso particular: el *Diccionario de colocaciones del español* (DiCE, Alonso Ramos 2002, 2004) y sus posibilidades de explotación del corpus incluido en él. Por último, esbozaremos algunas líneas futuras de esta investigación que apuntan hacia una concepción más abierta y dinámica del entorno que reúna diferentes recursos léxicos.

2. Desde el corpus como apoyo al diccionario hasta el diccionario como apoyo del corpus

En esta sección nos gustaría dar algunas pinceladas sobre la evolución del papel del corpus en relación con los diccionarios. En pocos años, el corpus está pasando de ser una fuente de la

que extraer ejemplos con los que dar testimonio de la información incluida en el diccionario hasta desplazar completamente al diccionario como un índice de la información incluida en él. En los dos extremos de esta tendencia se sitúan las dos vías que mencionamos antes: 1) PCED y 2) PDEC. Como veremos más tarde, estas dos estrategias no tienen por qué oponerse, pero, por el momento, mostremos someramente cómo se está dando la evolución de una estrategia a otra.

En la lexicografía actual, la estrategia PCED ya no se cuestiona. Como dijimos arriba, los diccionarios actuales necesitan corpus. Iztok y Krishnamurty (2007) son explícitos al respecto: "Publishers no longer gain an advantage over their competitors by using corpus data, but they can find themselves at a serious disadvantage by not doing so". La lexicografía española ya se ha sumado a esta corriente de explotar el corpus como apoyo al diccionario. Como pionero en este ámbito, destaca el diccionario publicado por la editorial SGEL a partir del corpus CUMBRE de 20 millones de palabras (Sánchez 2001). Más grande es el corpus empleado para la elaboración de *Redes* (Bosque 2004) que extrae su información de un corpus periodístico de 250 millones de palabras. Aquí, sin embargo, no nos interesa tanto el tamaño del corpus como el examen de cuál es el papel del corpus en relación con el diccionario.

El corpus puede desempeñar simplemente un papel de suministrador de ejemplos. De este modo, podríamos tomar un diccionario cualquiera y utilizar un corpus para buscar ejemplos que testimonien las acepciones ahí recogidas. Aquí el papel del corpus es solo un apoyo, aunque de esa tarea ya podría extraerse qué acepciones (entre otras cosas) no aparecen atestiguadas en el corpus, lo que empujaría a hacer limpieza en el diccionario. Sin embargo, en los diccionarios actuales el corpus está funcionando más como "director": si una acepción no aparece en el corpus, no entrará en el diccionario. Por esto, la lexicografía ha pasado de estar basada en el corpus (*corpus-based lexicography*) a ser dirigida por el corpus (*corpus-driven lexicography*). Un ejemplo de un diccionario español dirigido por el corpus puede ser el diccionario combinatorio *Redes.* Es a partir de la explotación del corpus desde donde se decide incluir una combinación como *chupetear vorazmente*, combinación marcada con el símbolo de "poco frecuente" pero atestiguada, frente a *comer* o *consumir vorazmente* que aparecen con el símbolo de "sumamente frecuente" y *leer vorazmente*, con el símbolo de "bastante frecuente". Sin un corpus, la primera combinación no hubiera entrado en el diccionario precisamente porque dada su poca frecuencia, no estaría en la disponibilidad léxica de los lexicógrafos redactores.

Ese papel de director otorgado al corpus puede ser contrarrestado por la introspección del lexicógrafo; en otras palabras, aunque el corpus dirige, no obliga. Siguiendo con el caso de *Redes*, el lexicógrafo bien puede optar por rechazar una combinación poco frecuente, bien puede optar por incluir combinaciones naturales pero que por puro azar no aparecen atestiguadas en el corpus, como es el caso de *ingerir vorazmente* (marcado como INDOC). En el caso de esta obra lexicográfica, han optado por combinar la *frecuencia*, concepto estadístico, con la *naturalidad*, concepto lingüístico (Bosque 2004: CLVIII). De hecho, han primado el concepto de naturalidad cuando optan por marcar con el símbolo de "sumamente frecuente" a combinaciones encontradas pocas veces pero que el lexicógrafo las percibe como sumamente naturales. Por esta razón, este diccionario se reclama como perteneciente a un tipo de *lingüística con corpus* (frente a *lingüística de corpus*): "una lingüística en la que el corpus está al servicio del investigador, de forma que los datos encontrados se filtran por su introspección, se evalúan y se completan con otros que el corpus no proporciona, pero que la introspección considera naturales" (Bosque 2004: CLIX).

Pues bien, frente a estas reservas hacia el corpus, planteadas desde un diccionario plenamente basado en el corpus, se empiezan a elevar voces que reducen el papel del diccionario al mínimo. Así, por ejemplo, Abaitua (2006), en su comunicación, avanza una

interesante idea en donde contrapone el *corpus como ayuda* a la producción de diccionarios al *corpus como diccionario*. En el primer caso, el diccionario constituye un producto acotado y, aunque esté basado en el corpus, solo representa una imagen estática del uso de la lengua[1]. Sin embargo, en el segundo caso, se trata de un proceso abierto, ininterrumpidamente, que dé muestras del uso cotidiano y de las innovaciones de la lengua, por lo que el diccionario pasa a diluirse en un corpus. Así, la lexicografía deja de ser un arte de hacer diccionarios a un arte de gestionar corpus. Abaitua (2006) predice que el corpus será el diccionario. En ese mismo trabajo aparece referenciado un artículo de Żmigrodzki (2005) que abunda en la misma línea. Para este autor, no tiene sentido publicar diccionarios en versión impresa y deberían aparecer como corpus explotados por algún buen programa de concordancias.

Hasta donde sabemos, quien ha avanzado más en la línea PDEC es Wanner (2006), que se ha centrado en la descripción de las colocaciones. Tras observar las varias limitaciones de los diccionarios de colocaciones en inglés, propone un corpus anotado como un diccionario de colocaciones. Entre otras ventajas, está la de ilustrar el uso de las colocaciones *in vivo* con corpus que se puede extender como el usuario desee, sin limitarse al par de ejemplos que el lexicógrafo habitualmente da, si es el caso, en un diccionario. Además, el corpus puede ser enriquecido con otros tipos de información semántica y sintáctica, de tal modo que el corpus se convierte en un recurso potente, capaz de nutrir otros recursos para el procesamiento automático de lengua natural (PLN). Para este autor, un corpus anotado con colocaciones y provisto de una interfaz con el usuario es una herramienta más adecuada que un diccionario convencional.

Aunque esta línea de pensamiento resulta atractiva, la estrategia PDEC tiene un inconveniente: su relación calidad/coste. El dilema sobre si resulta más costoso crear un diccionario apoyado en corpus o anotar un corpus con o sin diccionario no es fácil de resolver porque intervienen muchos factores; entre otros, si se utilizan o no técnicas automáticas para facilitar la explotación del corpus en el caso de la creación del diccionario. Simplificando un tanto el problema, podríamos decir que desde la estrategia PCED se abarcan menos datos pero se tratan con mayor calidad porque hay un filtrado del lexicógrafo, lo que resulta costoso; mientras que desde la estrategia PDEC se abarcan muchos datos pero la calidad es menor porque está más basado en técnicas automáticas con mayor o menor precisión y éxito. A pesar de los avances en PLN, anotar un corpus requiere un gran esfuerzo en tiempo y en recursos humanos y los resultados no son siempre óptimos, especialmente en lo que se refiere a la anotación semántica. En desambiguación automática de sentidos, la tarea consiste en etiquetar cada palabra del corpus con un sentido (Edmons y Kilgarriff 2002); para clasificar automáticamente colocaciones (Wanner et al. 2006), no solo hay que reconocer su patrón sintáctico sino también qué relación semántica se sostiene entre los elementos constituyentes de la colocación; para anotar los papeles semánticos en corpus como FrameNet (Ruppenhofer et al. 2006), la intervención humana ocupa un papel primordial y lo mismo, en corpus como Ancora (Aparicio et al. 2008) para el corpus catalán y español. Una ventaja a este respecto de

1 La idea del diccionario como una imagen estática del uso de la lengua recuerda las palabras, atribuidas a García Márquez y repetidas por muchos: "el diccionario es el cementerio en donde yacen las palabras muertas hasta que el hablante o el escritor las desentierra para resucitarlas y devolverlas a la vida". En el Prólogo del diccionario CLAVE, García Márquez, que se muestra como un gran admirador y usuario de los diccionarios, vuelve a manifestarse en la misma línea: "Los autores de los diccionarios las [las palabras] capturan demasiado tarde, las embalsaman por orden alfabético, y en muchos casos cuando ya no significan lo que pensaron sus inventores. En realidad, todo diccionario de la lengua empieza a desactualizarse desde antes de ser publicado".

la estrategia PCED es que desde esta perspectiva no es necesario enfrentarse a todo el corpus y de anotar todo lo que en él aparece, ya se trate de distinguir sentidos, ya de etiquetar papeles semánticos o clasificar colocaciones. Desde la estrategia PCED, la explotación del corpus es un requisito para el enriquecimiento del diccionario pero no se está obligado a cubrir todo el corpus. En el caso de un diccionario de colocaciones, como veremos más tarde, el corpus sirve para seleccionar las muestras de colocaciones que se quieren incluir en él. Con respecto a la desambiguación de sentidos, Kilgarriff (2005) se inclina también por la estrategia PCED: "What we would like is **some** corpus-based information about all dictionary senses, and it is immaterial if there are some corpus instances which do not contribute to any lexical entry".

La anotación del corpus es el núcleo de la estrategia PDEC, mientras que en la estrategia PCED, no se trata tanto de anotar el corpus como de explotarlo con el fin de apoyar la información incluida el diccionario. La explotación puede ser automática, semi-automática o manual pero va guiada por un objetivo léxico; es decir, el lexicógrafo que explota el corpus lo hace porque está buscando los sentidos de una determinada palabra, sus colocaciones o los papeles semánticos vinculados. No se enfrenta a un corpus corrido sino que va buscando información vinculada al lema de la entrada lexicográfica que está redactando. En cambio, en la estrategia PDEC, no hay objetivo léxico y el anotador debe empezar por la primera palabra y acabar por la última del texto incluido en el corpus. Obviamente, se pueden establecer estrategias de anotación como empezar etiquetando los predicados con sus argumentos, los verbos y sus funciones sintácticas o cualquier otra estrategia pero lo que queremos resaltar es que no hay el teleobjetivo preciso y específico desde el que el lexicógrafo busca la información sobre una unidad léxica dada.

Hasta ahora se han planteado las dos estrategias como una evolución de la PCED hacia la PDEC. Acabamos de ver que ambas presentan sus ventajas y sus inconvenientes. En la sección 4 plantearé un camino desde donde podremos combinar ambas estrategias, pero antes veamos algunas muestras que nos indican que quizás se trate de una falsa dicotomía. El concepto de diccionario está cambiando para convertirse en esa interfaz que reclamaba Wanner (2006). Si el diccionario es la interfaz que da acceso al corpus, ambos se funden en una nueva herramienta, como veremos a continuación.

3. Hacia una hibridación entre el corpus y el diccionario

Se podría proponer el neologismo "corpuscionario" para este nuevo híbrido que empieza a aparecer en los últimos tiempos. Quizás sea demasiado pronto para acuñar un nuevo término, pero sí que es ya el momento de llamar la atención sobre la desaparición de fronteras que se empieza a dar entre el corpus y el diccionario. No estamos pensando en nada esotérico sino simplemente en un corpus que se puede consultar alfabéticamente. Pensemos que una consulta así no se distingue en gran medida de un diccionario con información contextual. Las diferencias residen en qué información se incluya en el diccionario y qué anotación reciba el corpus, pero en sí, la diferencia no será conceptual sino de plataforma sobre la que mostrar la información. Uno de los grandes impulsores de la vinculación entre corpus y diccionario como fue Sinclair pensaba que el diccionario no es otra cosa que un comentario sobre los ejemplos. En su conferencia, *The Dictionary of the Future*, Sinclair concibe el diccionario explícitamente como una interfaz de acceso al corpus: "a device through which the user will observe the living language" (Sinclair 1987).

Sin querer hacer un repaso exhaustivo, nos gustaría describir aquí algunas muestras de estos híbridos. Hemos escogido cuatro: la primera tiene principalmente aplicaciones en PLN, aunque de ese léxico puede derivarse también un diccionario para humanos; la segunda está construida esencialmente sobre la estadística de un corpus; la tercera consiste en una

reutilización de un corpus paralelo aunque complementado con otros repertorios lexicográficos; y la cuarta es una herramienta pensada para la ayuda a la producción de diccionarios que pasa a convertirse en un diccionario.

3.1. Léxicos verbales y Corpus de Ancora

Empecemos por mostrar cómo el diccionario y el corpus pueden retroalimentarse. Los léxicos verbales del catalán y del español vinculados al corpus AnCora (Aparicio et al. 2008a, 2008b) se obtienen del corpus y son, a su vez, utilizados para anotar semánticamente el corpus. El corpus AnCora, además de anotación morfológica y sintáctica, recibe también anotación semántica con los papeles semánticos asignados a los argumentos de los predicados verbales que aparecen en el corpus[2]. Precisamente para la anotación semántica, es crucial el papel de los léxicos verbales. Estos se crean manualmente a partir del corpus analizado sintácticamente. A cada UL verbal se le asigna una clase semántica, la estructura argumental y las alternancias de diátesis. Con esta información en el léxico, se extraen algunas reglas de proyección que sirven para anotar semánticamente el corpus de tal modo que se asigne a un predicado verbal dado con una clase semántica dada, una estructura argumental específica y los papeles semánticos asignados a los argumentos expresados en el corpus. A modo de ilustración, reproducimos aquí parte de la entrada del verbo catalán *abonar* y un ejemplo de una frase con ese verbo en el corpus anotado.

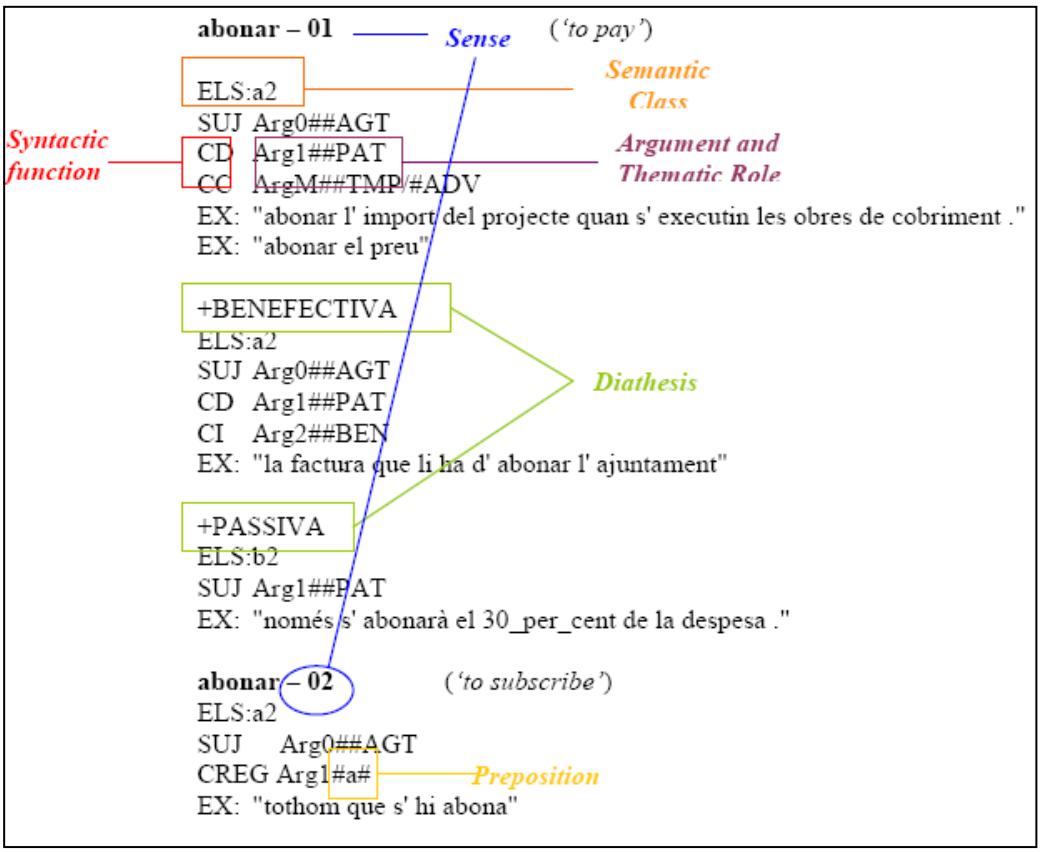

Fig. 1. Entrada del verbo catalán *abonar*

2 Tanto el corpus como los léxicos asociados pueden ser consultados en la siguiente página web: http://clic.ub.edu/ancora/.

El president sosté que els va abonar amb fons reservats

S

- **sn SUJ** ARG0 AGT
  - spec
    - d ***El*** el
  - grup.nom
    - n ***president*** president
- **grup.verb**
  - v ***sosté*** sostenir a2
- **S CD** ARG1 PAT
  - conj
    - c ***que*** que
  - **sn SUJ** ARG0 AGT [... ø ...]
  - **sn CD** ARG1 PAT
    - grup.nom
      - p ***els*** ell
  - **grup.verb**
    - v ***va*** anar
    - v **abonar** abonar a2
  - **sp CC** ARGM ADV
    - prep
      - s ***amb*** amb
    - sn
      - grup.nom
        - n ***fons*** fon
        - S
          - participi
            - a ***reservats*** reservat

f . .

Fig. 2 Extracto del corpus anotado con el verbo *abonar*

Una vez que el corpus esté anotado, puede decirse que el léxico no es más que la interfaz que permite consultar alfabéticamente el corpus y reúne las generalizaciones que se extraen a partir de la anotación de todas las instanciaciones de una UL verbal.

3.2. El corpus WortSchatz de la Universidad de Lepizig

Una muestra interesante de combinación de corpus y diccionario es *WortSchatz*. Esta herramienta, compilada por la Universidad de Leipzig, es concebida como un corpus, pero de hecho, el resultado de su consulta ofrece datos como los que ofrece un diccionario con información extraída del corpus. Se puede consultar tecleando una palabra como en un diccionario ordinario. Es accesible libremente a través de internet[3] y se alimenta diariamente, puesto que presenta cada día las palabras nuevas más frecuentes. Aunque pensada previamente para el alemán, hoy se puede consultar en 48 lenguas, aunque no todas con el

3 En el enlace siguiente se puede consultar: http://wortschatz.uni-leipzig.de/. Para profundizar en cómo se ha compilado el corpus por un equipo de procesamiento de lengua natural de la Universidad de Leipzig, vid. Biemann et al. (2004, 2007) y Richter et al. (2006).

mismo acabado. Su facilidad de acceso la convierte en una herramienta especialmente útil para los aprendices de alemán como lengua extranjera. En el trabajo de Oster (2009), se puede consultar una cibertarea apoyada en esta herramienta para que los aprendices de alemán conozcan la palabra *Karaffe*. El corpus proporciona cuatro tipos principales de información: una descripción, palabras relacionadas, coocurrentes significativos y contextos. A continuación ponemos un extracto de la información que ofrece para el sustantivo *Karaffe*.

Beschreibung 'descripción'**:**	bauchige Flasche 'botella de cuerpo ancho' geschliffene Glasflasche 'botella de cristal tallado'
Sachgebiet 'dominio'	Behälter 'contenedor'
Morphologie:	kar|aff|e
Grammatikangaben 'información gramatical'**:**	Wortart 'clase de palabras': Substantiv
	Geschlecht ,género': weiblich ,femenino Flexion: die Karaffe, der Karaffe, der Karaffe, die Karaffe die Karaffen, der Karaffen, den Karaffen, die Karaffen

Relationen zu anderen Wörtern 'relación con otras palabras'**:**

- Synonyme: Behälter, Flasche, Gefäß, Kanne, Krug

Links zu anderen Wörtern 'vínculos con otras palabras'**:**

- ~~Grundform 'forma básica' :~~ Karaffe
- ist ein(e) 'es un' Gefäß, Glasflasche
- Synonym von: Wasserflasche
- Form(en): Karaffe, Karaffen

Beispiel(e) 'ejemplo(s)'**:**

Also gießen Sie die Flasche in eine **Karaffe**?

Dann entwickeln sich in der **Karaffe** Aromen von Schlehe, roten Johannisbeeren, einem Hauch von Paprika,

Datteln und Zimt.

Er bestellt eine **Karaffe** Rotwein, ein Baguette mit etwas Käse, schließlich ist man in Paris.

Signifikante Kookkurrenzen für Karaffe 'coocurrencias significativas de Karaffe'**:**

Wein (54), Wasser (45), eine (36), Johannisbeeren (35), Aromen (29), Rotwein (24), man (23), bestellen (21),

Signifikante linke Nachbarn von Karaffe 'vecinos a la izquierda significativos'**:**

eine (116), einer (17)

Signifikante rechte Nachbarn von Karaffe 'vecinos a la derecha significativos'**:**

Rotwein (36), Wein (29), mit (24), Wasser (22), etwas (17)

Fig. 3. Información sobre *Karaffe* en WortSchatz

A partir de esta información, Oster (2009) prepara una ficha para que los aprendices rellenen la siguiente información con respecto a las coocurrencias:

Ficha

(...)

c) Kollokationen / *colocaciones***:**

Analiza los "vecinos a la derecha" (Rechte Nachbarn) y una página de ejemplos adicionales

(weitere Beispiele). Apunta toda la información que encuentres sobre:

- Was ist drin? / *¿Qué contiene?* (sustantivos): [Wein, Chianti, Wasser, Rotwein / *vino, chianti, agua, vino tinto*]
- Material (adjetivos): [silbern, gläsern / *de plata, de cristal*]
- Wofür? / *¿Para qué sirve?* (verbos): [dekantieren / *decantar*]

Fig. 4. Ficha didáctica sobre *Karaffe* (Oster 2009**)**

No está claro, con todo, cómo el aprendiz obtiene la coocurrencia de verbos como *decantar* a partir de la información ofrecida en el *WortSchatz*. Por lo demás, mucha de la información ahí ofrecida se basa únicamente en la estadística y es de poco interés desde un punto de vista lexicográfico. Pensemos, por ejemplo, en los "vecinos a la izquierda": no hay ningún interés lexicográfico en indicar que un nombre coocurre con un artículo indeterminado.

3.3. Diccionario CLUVI

Otra muestra del diccionario como interfaz de acceso al corpus puede ser el *Diccionario CLUVI inglés-galego* (Gómez Guinovart 2008). Este diccionario está basado en un corpus paralelo alineado. Tanto el corpus como el diccionario son disponibles libremente en la web[4]. En Gómez Guinovart et al. (2008), los autores explican cómo fue el proceso de construcción del corpus alineado y cómo derivaron, en primera instancia, un diccionario bilingüe probabilístico al que siguió una revisión manual para mejorar su precisión. En la segunda

4 Puede consultarse en la dirección siguiente: http://sli.uvigo.es/CLIG/

edición, el lemario recoge todos los lemas que aparecen un mínimo de tres veces en el corpus, además de un conjunto de lemas que consideraron interesantes, a pesar de ser menos frecuentes en los textos. La microestructura consta principalmente de los ejemplos de contextualización en L1 y en L2. Por lo tanto, el diccionario constituye una interfaz de acceso a un corpus paralelo puesto que muestra solo los equivalentes de traducción que se encuentran en el corpus. Así, por ejemplo, el adjetivo *heavy* ofrece solo dos equivalentes en gallego (*pesado* e *forte*) a pesar de las otras muchas posibilidades de traducción de ese adjetivo, precisamente porque es muy productivo en colocaciones en inglés. Si un equivalente no aparece en el corpus, no tendrás la correspondencia en el diccionario bilingüe[5]. Otro posible problema que puede encontrarse en la correspondencia entre unidades de traducción en un corpus dado es la derivación de equivalencias ocasionales al estatuto de equivalencias generales; es decir, un traductor en un momento dado puede optar por una formulación como una estrategia de traducción sin que pueda considerarse que sean equivalentes fuera de esa ocasión. En la documentación asociada al *Diccionario CLUVI* no hemos encontrado cuántas veces debe aparecer un equivalente de traducción en el corpus para poder ser derivado al diccionario. Así, por ejemplo, el hecho de que el nombre inglés *heat* aparezca traducido como *vehemencia* en ese corpus no es suficiente para derivar al diccionario bilingüe la equivalencia inglés-gallego: *heat-vehemencia*. Es cierto que ese nombre en inglés tiene un sentido 'sentimiento fuerte de rabia o excitación' y que en ciertos contextos puede traducirse como *vehemencia*, pero en un diccionario bilingüe, la equivalencia más generalizada para ese sentido debería ser *acaloramento*[6]. En cambio, del diccionario no se puede inferir que haya más posibilidades de traducir ese sentido puesto que la única información que aparece para describir esa equivalencia es el extracto del corpus paralelo:

EN *He turned on her cheek the* ***heat*** *of love, its horror, its cruelty, its unscrupulosity.*

GL *Espetoulle na cara a* ***vehemencia*** *do amor, o seu horror, a súa crueldade, a súa falta de escrúpulos.*

3.4. Diccionario de colocaciones automático GDEX

Si empezábamos esta sección con el visionario Sinclair, nos gustaría también acabar con él. Según Moon (2008), en el momento de redactar el *Collins Cobuild* se consideró incluir las secuencias de ejemplos antes de las definiciones con el objetivo de mostrar las pruebas antes de la explicación y así permitir a los usuarios localizar el significado heurísticamente. Más tarde, Sinclair planeó un diccionario de colocaciones, estructurado simplemente alrededor de las concordancias, que nunca fue completado. Esta idea hubiera sido el prototipo de diccionario en el que las palabras no son más que los puntos de acceso al corpus, a los textos que muestran el significado. Algo similar es la herramienta ofrecida por Kilgarriff et al. (2008), basada en la tecnología Sketch Engine (Kilgarriff et al. 2004). Con un análisis sintáctico superficial del corpus, su herramienta ofrece palabras que coocurren con la palabra consultada y se da además la relación sintáctica que las une. A continuación ponemos un extracto de los colocativos verbales de la entrada para *opinion*, en inglés[7]

---

5 Esta objeción de la dependencia del corpus no está vinculada al proceso de hibridación del que venimos hablando sino a toda empresa lexicográfica que dependa exclusivamente del corpus.

6 En el TILG (http://www4.usc.es/TILG/) pueden encontrarse varios ejemplos de *acaloramento* que responden a ese sentido de *heat*.

7 Un prototipo de este diccionario es disponible en http://forbetterenglish.com

object_of	
express	No one had ever seen Pike *express an opinion* about anything.
voice	Try to get teachers to *voice their opinions* on important subjects.
form	Firstly, the role of the news media in *forming public opinion* is very important.
divide	In fact, the general tide of expert *opinion* is deeply *divided*.
seek	Still, she was pleased he had *sought her opinion*.
change	At the very beginning of the play Shakespeare demonstrated how easily the people *changed their personal opinions*.

Fig 5. Extracto del Diccionario GDEX

Aunque Kilgarriff no ha concebido su herramienta como sustitución al diccionario, podríamos pensar en la línea de Sinclair que esa información es suficiente y que el diccionario es solo el punto de acceso al corpus. Sin embargo, seguimos pensando que el corpus no reemplaza al diccionario porque este proporciona la reflexión analítica de los datos. En la sección siguiente, presentaremos un modo de poder acceder al corpus a través de un diccionario pero un corpus enriquecido por la información del diccionario.

4. El corpus en el diccionario: el caso del *Diccionario de colocaciones del español*

Hasta aquí hemos mostrado el papel preponderante que está tomando el corpus hasta llegar a desplazar o sustituir al diccionario. Ahora vamos a proponer otra manera de hacer interactuar corpus y diccionario. Con la metodología actual de compilar diccionarios inductivamente a partir de corpus, el propio diccionario puede convertirse en el "contenedor" del corpus. 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. Este es el caso del *Diccionario de colocaciones del español* (DiCE, Alonso Ramos 2002, 2004). Por tanto, el DiCE se enmarcaría en la estrategia PCED. Se basa en un corpus que resulta enriquecido por la información que le aporta el diccionario. Los defensores de la estrategia PDEC podrán argumentar que un corpus anotado también está enriquecido, pero la contrargumentación que se puede hacer desde el DiCE es que incluso un corpus sin anotar, al estar asociado a la información del diccionario, ya está enriquecido. La estrategia PDEC suele dejar que el corpus "hable solo" y se confía mucho en la inferencia que tiene que hacer el usuario. Sin embargo, pensamos que al usuario, especialmente al aprendiz, de poco le sirve que se le ofrezca una colocación si no está desambiguados semánticamente ni la base ni el colocativo (como la demo del GDEX). En contraste, nosotros pensamos que la información debe ser lo más explícita posible y por ello, el diccionario no solo debe dar ejemplos sino que tiene que comentarlos o explicarlos, como decía Sinclair. Antes de profundizar más en cómo poder explotar el corpus contenido en el propio diccionario, necesitamos presentar brevemente el DiCE y mostrar qué información se ofrece.

4.1. Breve presentación del DiCE

El DiCE ha sido concebido como una base de datos, que se puede consultar en la web[8]. Se caracteriza por los siguientes tres rasgos: (1) cada colocación recibe una descripción semántica y sintáctica; (2) cada colocación es atestiguada con varios ejemplos, la mayoría extraídos del *Corpus de referencia del español actual* (CREA); y (3) está asociado con un módulo didáctico. Examinemos brevemente cada uno de estos aspectos.

Para describir las colocaciones, usamos las *funciones léxicas* (FLs), la herramienta de la Lexicología explicativa y combinatoria (Mel'čuk *et al.* 1995), que ha sido ampliamente rodada en diferentes proyectos lexicográficos (para el francés Mel'čuk *et al.* 1984/1999 y el más reciente Mel'čuk y Polguère 2007). Una FL codifica la relación entre dos unidades léxicas de las cuales 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 adjetivo-nombre: *honda pena*, *terrible vergüenza*, y *ferviente admiración*. Los tres adjetivos son seleccionados para expresar, en combinación con el nombre correspondiente, el mismo significado, aproximadamente 'intenso'. Las FLs son el mejor instrumento para describir las colocaciones porque satisfacen tres requisitos indispensables para un recurso léxico operativo: 1) proporcionan el significado de la colocación; 2) describen la sintaxis y la estructura argumental de la colocación; y 3) codifican la dependencia funcional del colocativo en relación con la base. Para facilitar el uso de las FLs, en el DiCE se ha optado por usar *glosas* en lengua natural que codifican el significado de las colocaciones. La glosa puede ser considerada como la traducción de una FL a una metalengua natural (Alonso Ramos 2006a). Por lo tanto, los usuarios pueden acceder al colocativo por medio de la glosa en lugar de la FL. Así, por ejemplo, la glosa para la FL Magn es 'intenso'.

Con respecto al segundo aspecto, como hemos dicho, todas las colocaciones están apoyadas en ejemplos extraídos del corpus. Ahora bien, aunque el DiCE se compila inductivamente a partir del corpus, las FLs constituyen una plantilla que guía la búsqueda de colocaciones en el corpus. El corpus por tanto es filtrado desde el principio con búsquedas específicas. Así, por ejemplo, a la hora de buscar las colocaciones de un nombre como *opinión*, el análisis semántico junto con la plantilla de las FFLL lleva al lexicógrafo a buscar colocativos específicos. No buscará, por ejemplo, un valor de la FL Magn porque el significado de ese nombre no es compatible con la intensificación, pero sí buscará:

1) adjetivos que expresan cuántas personas coinciden en la opinión: *mayoritaria*, *generalizada*, *compartida*, *personal*;

2) adjetivos que caracterizan si el contenido de la opinión es positivo o negativo: *buena*, *mala*, *contraria*, *favorable*;

3) verbos que toman el nombre como objeto para expresar que se tiene una opinión: *tener*, *sostener*;

4) verbos que toman el nombre como objeto para expresar que se manifiesta una opinión: *expresar*, *dar*, *ofrecer*;

5) verbos que toman el nombre como objeto para expresar que el contenido de la opinión varía: *formarse*, *cambiar*;

6) verbos que toman el nombre como sujeto para expresar que hay muchos con la misma opinión: *extenderse*

7) etc.

Puesto que cada colocativo encontrado en el corpus es recogido en la base de datos con su contexto, el DiCE puede ser usado como un corpus de colocaciones, que puede ser de gran

8 La dirección es: http/www.dicesp.com. Actualmente, el DiCE en la web está en fase de remodelación que esperamos termine en la próxima primavera.

utilidad para el usuario. A modo de ilutración, la información para una colocación como *dar una opinión* aparecería así:

VERBO + OPINIÓN

'expresar ~ '$_{\text{FL Caus1Manif}}$ = dar [ART ~ sobre Z/ a W] *No soy capaz de* dar una opinión*, cuando me la piden me pongo muy nervioso; Anteriormente ya* di mi opinión sobre *este producto; me pidieron que viera a unos cinco jugadores y les* di mi opinión.

En cuanto al tercer aspecto, el módulo didáctico es todavía muy preliminar, pero, incluso así, la descripción semántica y sintáctica de las colocaciones en el DiCE permite ya una explotación interesante del corpus para el aprendiz de español.

4.2. Posibilidades de explotación del corpus de colocaciones en el DiCE

El corpus contenido en el DiCE presenta unas características especiales que lo hacen particular. Puesto que la selección de ejemplos es manual[9], el corpus no pasa por ninguna fase de anotación antes de pasar a ser registrado en el DiCE. El lexicógrafo selecciona los ejemplos con el "teleobjetivo" del que hablábamos arriba, lo que da una mayor calidad al corpus que ejemplos tomados completamente al azar. Sin lugar a dudas, herramientas como el GDEX (Kilgarriff et al. 2008) que facilitan la tarea al lexicógrafo de seleccionar el mejor ejemplo son bienvenidas, pero esto no obsta para que siga primando el criterio de lexicógrafo y no el criterio del dado del azar como en una herramienta basada exclusivamente en corpus y no en diccionario.

Otra característica relevante del corpus en el DiCE es que aunque no está anotado está enriquecido. Como vimos arriba, 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, del extracto del corpus *Anteriormente ya* di mi opinión sobre *este producto*, al ir asociado 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". A partir de aquí no sería demasiado difícil etiquetar el papel semántico de los argumentos que aparecen en los ejemplos, con lo que se obtendría un corpus de colocaciones con los papeles semánticos anotados (Prieto González 2008).

Llegados a este punto, podemos plantearnos hasta qué punto la distinción establecida desde el principio entre la estrategia PCED y la PDEC no es una falsa dicotomía y quizás solo se trate de dos vías de llegar a un mismo resultado. Así, si se anotara un corpus de colocaciones, en la línea propuesta por Wanner (2006) y paralelamente, se utilizara ese mismo corpus para vincularlo al DiCE, se llegaría a la misma información. La diferencia que es en el primer caso se es esclavo del corpus, mientras que en el segundo no, porque el lexicógrafo siempre pueda incluir colocaciones que no están en el corpus de trabajo.

---

9 Ha habido distintos experimentos de identificación automática de colocaciones, pero con un éxito limitado (vid. Heid y Weller 2008, Villavicencio et al. 2005). Con todo, hay que señalar dos factores:1) las diferentes interpretaciones que se tiene de lo que es una colocación hacen que el éxito o el fracaso del experimento sea medido de distinta manera; 2) la extracción de las colocaciones y la clasificación semántica son dos tareas distintas; los que se ocupan de la primera se basan esencialmente en estadística, mientas que para la segunda, hace falta conocimiento lingüístico (Wanner et al. 2006).

Para poder explotar el corpus contenido en el DiCE, especialmente con fines didácticos, es necesario añadir una herramienta de búsqueda que navegue sobre el corpus. Esta herramienta funcionaría de un modo parecido a un simple programa de concordancias, pero capaz de buscar ejemplos vinculados al mismo lema, a pesar de que el corpus no esté directamente lematizado. Los usuarios, aprendices de ELE y también usuarios nativos con dudas específicas en el momento de la redacción, usarían esta herramienta de navegación cuando no están necesariamente interesados en consultar la entrada lexicográfica de la base de la colocación porque solo quieren aclarar una duda específica. Por ejemplo, si un aprendiz quiere saber si el nombre *opinión* tiene o no un determinante cuando va con el verbo *dar*, en lugar de ir a la entrada del nombre y recorrer toda la información, sería más rápido y eficaz lanzar la herramienta de búsqueda que navegue sobre el corpus incluido en el diccionario. La herramienta busca la coocurrencia entre *dar* y *opinión*, y puesto que los ejemplos del valor *dar* son agrupados por el lema, el corpus no necesita estar lematizado.

Un ejemplo más interesante se da cuando la búsqueda pedida se corresponde con dos descripciones. Por ejemplo, un usuario puede preguntarse si el nombre *respeto* se combina con el verbo *tener*. La herramienta de búsqueda devolvería los ejemplos separados en dos grupos clasificados según la FL correspondiente: *tener respeto a alguien*, codificado por la FL $Oper_1$ que representa los verbos soporte o de apoyo, y *tener el respeto de alguien*, codificado por la misma FL pero con otro subíndice actancial, $Oper_2$ para dar cuenta de la conversión de acantes. Mientras la primera colocación es una paráfrasis del verbo *respetar*, la segunda sería parafraseada por la pasiva *ser respetado*. Los datos en el DiCE se muestran así:

‘sentir ~ ’$_{\text{FL Oper1}}$ = tener [~ a Y] |

Sus contemporáneos le tienen un gran respeto; Yo a su hermano le tengo cariño y respeto; ¿Quién me habría tenido el menor respeto si yo hubiera cambiado?

‘ser objeto de ~ ’$_{\text{FL Oper2}}$ = tener [ART ~] | artículo obligatorio; expresión obligatoria del actante X

Tiene el respeto de todos sus contemporáneos; Tenía la admiración y el respeto de los que le escuchaban

Como vemos, el corpus dentro del DiCE es un corpus enriquecido. Ahora bien, como todo corpus “contenido” en otra herramienta presenta la desventaja de ser limitado. El corpus puede presentar lagunas por distintas razones: 1) por puro azar, una colocación dada no apareció en el corpus explotado; 2) por negligencia/error del lexicógrafo, una colocación puede haber pasado desapercibida; 3) por la reciente aparición de un colocación en la lengua; etc. Dadas estas limitaciones, creemos que es importante que desde la propia herramienta se dé acceso al corpus libre o al corpus exterior a la propia herramienta. Especialmente con fines didácticos, es importante subrayar la idea de que el diccionario no tiene el poder sagrado que atribuyen algunos usuarios ingenuos (“si está en el diccionario, está bien; si no, está mal”). Especialmente en el tema de las colocaciones, en donde 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’. Por esta razón, es importante que desde el recurso léxico se le dé entrada al corpus libre y el usuario pueda

consultar la web desde un motor como Google[10] o consultar el CREA. En este caso concreto, la consulta desde Google de la combinación "admirador empedernido", le daría 195 ocurrencias, lo que no es muy alto. Desde el CREA, la consulta "admirador dist/5 empedernido" le 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. Un buen ejemplo en esta línea nos lo ofrece Milton (2006) que sugiere añadir a su herramienta una caja de diálogo con el buscador de Google en donde advierte a los aprendices que pueden encontrar ejemplos no estándar que no son aceptables. La caja facilita las consultas a los usuarios, evitándoles la necesidad de aprender a usar asteriscos o a usar los operadores booleanos típicos de los lenguajes de búsqueda. En la última sección, abogamos por una apertura de los recursos.

5. Por un entorno más completo y dinámico

El usuario actual de diccionarios, especialmente el aprendiz avanzado de lenguas, está habituado a consultar diccionarios web y consultar Google como corpus. Ooi (2008) es claro al respecto: "Nowadays, the user is not only encouraged to combine the strengths of multiple dictionaries and online encyclopedias but also to sift through more information in order to get to the required definition and meaning". Por esta razón, pensamos que los nuevos recursos léxicos deben ser concebidos en la línea de los entornos de trabajo de los traductores, a los que se puede considera usuarios expertos (Rogers y Ahmad 1998). Se trata de diseñar entornos en donde se combinen distintos recursos: diccionarios monolingües, diccionarios bilingües, tesauros, enciclopedias, corpus procesado y corpus libre. Si ese entorno tiene como destinatarios también a aprendices de lengua, hay que añadir un módulo didáctico que se apoye en todos esos recursos (Alonso Ramos 2006b). Aunque hemos puesto especial atención a los diccionarios de colocaciones, también nos gustaría señalar la importancia de combinar diccionarios monolingües con el diccionario de colocaciones: el *Oxford Phrasebuilder Genie* es una prueba de ello. Obviamente, la vinculación entre diccionarios de colocaciones de distintas lenguas también mejoraría la herramienta porque permitiría buscar equivalencias paralelas entre colocaciones. Por ejemplo, permitiría evitar la búsqueda en un bilingüe inglés-español de cómo se traduce al español *heavy* cuando va con *smoker* porque esta información vendría codificada en cada diccionario de colocaciones monolingüe correspondiente.

Una muestra de cómo podría ser ese entorno se puede encontrar en la *Base lexicale du français* (Verlinde et al. 2006), accesible en la web (http://ilt.kuleuven.be/blf/), en donde la combinación de recursos léxicos, corpus y actividades didácticas lo convierte en una herramienta deseable para el español. El usuario puede optar por utilizar un recurso u otro, pero lo que parece evidente es que la combinación de diccionarios y corpus es indispensable, como hemos querido mostrar a lo largo de este trabajo.

10 De hecho la investigación sobre la web como corpus aumenta cada vez más. Vid. Kilgarriff y Grefenstette (2003).

Referencias bibliográficas

Abaitua, J. (2006): "Taxonomías y ontologías para la gestión de recursos lexicográficos", en Atti del Convegno Internazionale *Glossari, dizionari, corpora: Lessicologia e lessicografia delle lingue europee*. Gargnano del Garda, Italia (25-27 mayo 2006), Università degli Studi di Milano.

Alonso Ramos, M. (2002): "Un vacío en la enseñanza del léxico del español como lengua extranjera", en A. Braasch and C. Povlsen (eds.), *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002,* volume II, Copenhague, CST, 551-561.

Alonso Ramos, M. (2004): "Elaboración del Diccionario de colocaciones en español y sus aplicaciones", en P. Bataner and J. de Cesaris (eds.), *De Lexicographia. Actes del I Symposium internacional de Lexicografia*, Barcelona, IULA-Edicions Petició, 149-162.

Alonso Ramos, M. (2006a): "Glosas para las colocaciones en el *Diccionario de colocaciones del español*", en Alonso Ramos, M. (ed.). *Diccionarios y fraseología* (*Anexos de Revista de Lexicografía*, 3), A Coruña, Universidade da Coruña, 59-88.

Alonso Ramos, M. (2006b): "Towards a dynamic way to learn collocations in a second language", en Corino, E.; Marello, C.; Onesti, C. (eds.). *Proceedings of the Twelfth EURALEX International Congress*. Torino: Accademia della Crusca, Università di Torino, Edizioni dell'Orso Alessandria. 909-923.

Aparicio, J., M. Taulé, M.A. Martí (2008a): "AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora". *Proceedings of 6th International Conference on Language Resources and Evaluation*. Marrakesh (Morocco).

Aparicio, J., M. Taulé, M.A. Martí (2008b): "AnCora-Verb: Two Large-scale Verbal Lexicons for Catalan and Spanish", *Proceedings of XII Euralex*, Barcelona (Spain).

Biemann, C., S. Bordag, G. Heyer, U. Quasthoff, C. Wolff (2004): "Language independent Methods for Compiling Monolingual Lexical Data", en *Computational Linguistics and Intelligent Text Processing* (Proceedings of CicLING 2004), Springer Lectures Notes in Computer Science vol. 2945. Seoul, South Korea, p.217-228

Biemann, C, G. Heyer, U. Quasthoff, M. Richter (2007): "The Leipzig Corpora Collection: Monolingual Corpora of Standar Size", en *Proceedings of Corpus Linguistic 2007*, Birmingham, UK.

Bosque, I. (dir.) (2004): *Redes. Diccionario combinatorio del español contemporáneo*, Madrid, SM.

Edmonds, P, A. Kilgarriff (2002): "Introduction to the Special Issue on Evaluating Word Sense Disambiguation Systems". *Journal of Natural Language Engineering* 8 (4).

Gómez Guinovart, X. (coord) (2008). Diccionario CLUVI inglés-galego (2ªedición) [http://sli.uvigo.es/CLIG/]

Gómez Guinovart, X, E. Díaz Rodríguez, A. Álvarez Lugrís (2008): "Aplicacións da lexicografía bilingüe baseada en córpora na elaboración do Dicionario CLUVI inglés-galego". *Viceversa: Revista Galega de Traducción*, 14.

Heid, U., M. Weller (2008): "Tools for Collocation Extraction: Preferences for Active vs. Passive", *Proceedings of the Sixth International Language Resources and Evaluation* LREC'08, Marrakech, Morocco.

Iztok K., R. Krishnamurthy (2007): " A New Venture in Corpus-Based Lexicography: towards a Dictionary of Academic English", en *Proceedings of Corpus Linguistics 2007*.

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

Kilgarriff, A., G. Grefenstette (2003): "Introduction to the Special Issue on Web as Corpus". *Computational Linguistics* 29 (3),
Kilgarriff, A., P. Rychly, P. Smrz, D. Tugwell (2004): "The Sketch Engine". *Proc. EURALEX*, Lorient, France, 105-116.
Kilgarriff, M. Husák, K. McAdam, M. Rundell, P Rychlý (2008): "GDEX: Automatically finding good dictionary examples in a corpus". Proc EURALEX, Barcelona, Spain.
Krishnamurty, R. (2008), "Corpus-driven Lexicography", *International Journal of Lexicography*, vol. 21, 3, 231-242
Mel'čuk, I.A. *et al*. (1984-1999): *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques* I-IV. Montréal: Les Presses de l'Université de Montréal.
Mel'čuk, I.; Clas, A.; Polguère, A. (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", en K. Hyland and F. Hyland (eds.) *Feedback in Second Language Writing: Contexts and Issues*, Cambridge, Cambridge University Press.
Moon, R. (2008), "Sinclair, Phraseology, and Lexicography", *International Journal of Lexicography*, vol. 21, 3, 243-254.
Ooi, V.B.Y (2008): "The Lexis of Electronic Gaming on the Web. A Sinclairian Approach", *International Journal of Lexicography*, vol. 21, 3, 311-323.
Oster, U. (2009): "La adquisición de vocabulario en una lengua extranjera: de la teoría a la aplicación didáctica", *Porta linguarum* 11, 33-50.
Prieto González, S. (2008): "Inclusión de los papeles semánticos de FrameNet en DiCE", *Proceedings of Euralex*, Barcelona.
Rogers, M., K. Ahmad (1998): "The Translator and The Dictionary: Beyond Words?", en B. T. S. Atkins (ed.) *Using dictionaries*, Niemeyer, Tübingen, 193-204.
Richter, M., Quasthoff, U., Hallsteinsdóttir, E., Biemann, C (2006): "Exploiting the Leipzig Corpora Collection", en *Proceedings of the IS-LTC 2006.* Ljubljana, Slovenia.
Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C. R. and Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice*. http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=126
Sánchez, A. (dir.) (2001). *Gran diccionario de uso del español basado en el Corpus lingüístico CUMBRE*. Madrid, Sociedad General Española de Librería.
Verlinde, S, T. Selva, J. Binon (2006): « The Base lexiclae du français (BLF : A multifunctional online database for learners of french », Proceedings of XII Euralex, Torino, 471-483.
Sinclair, J. M. (1987). *The Dictionary of the Future*. Collins English Dictionary Annual Lecture. University of Strathclyde, 6 May 1987.
Villvicencio, A., F. Bond, A. Korhonen, D. McCarthy (eds.) (2005). Special issue on Multiword Expression, *Computer Speech & Language*Volume 19, Issue 4.
Wanner, L. (2006): "¿El corpus como un Diccionario de colocaciones?", en Alonso Ramos, M. (ed.). *Diccionarios y fraseología* (*Anexos de Revista de Lexicografía*, 3), A Coruña: Universidade da Coruña.
Wanner, L., B. Bohnet, M. Giereth, y V. Vidal (2006): „Making Sense of Collocations". *Computer Speech & Language* 20(4), 609-.624.

Żmigrodzki, P. (2005): “Dictionary as a Text Corpus - Text Corpus as a Dictionary”. *Perspectives of Scholarly Lexicography in Poland*.