MONOGRÁFICO

Estudios de **Traducción**
ISSN: 2174-047X

EDICIONES COMPLUTENSE

https://dx.doi.org/10.5209/estr.87191

PLATCOL, Plataforma Multilingüe de Diccionarios de Colocaciones: el caso del chino[1]

José Manuel Pazos Bretaña[2], Adriane Orenha Ottaiano[3], Zhongmei Xiong[4]

Recibido: 27 de febrero de 2023 / Aceptado: 2 de junio de 2023

Resumen. El objetivo de esta contribución es realizar algunas observaciones sobre el procesamiento de las colocaciones extraídas de la lengua china, así como discutir los problemas que hemos observado al trabajar con esta lengua en la Plataforma Multilingüe de Diccionarios de Colocaciones (*PLATCOL*). PLATCOL incluirá colocaciones en inglés, portugués, español, francés y chino (Orenha-Ottaiano et al. 2021) y forma parte del proyecto *A phraseographical methodology and model for an Online Corpus-Based Multilingual Collocations Dictionary Platform* (Proceso *FAPESP* 2020/01783-2). En la plataforma se ha seguido una metodología unificada para obtener los datos que poblarán las entradas. Esta metodología que funciona con razonable eficacia en las demás lenguas –aunque requiere una fase supervisada de corrección y validación– conlleva un esfuerzo suplementario en el caso de la lengua china donde, por ejemplo, discrepancias en la asignación de categorías gramaticales pueden afectar a la eficacia del método a la hora de extraer candidatos.
Palabras clave: colocaciones; fraseografía; lexicografía electrónica; chino mandarín

[en] PLATCOL, Multilingual Platform for Collocation Dictionaries: the case of Chinese

Abstract. The aim of this contribution is to make some observations on the processing of collocations extracted from the Chinese language, as well as to discuss the problems we have encountered when working with this language in the Multilingual Platform for Collocational Dictionaries (PLATCOL). PLATCOL will include collocations in English, Portuguese, Spanish, French and Chinese (Orenha-Ottaiano et al. 2021) and is part of the project *A phraseographical methodology and model for an Online Corpus-Based Multilingual Collocations Dictionary Platform* (Proceso *FAPESP* 2020/01783-2). A unified methodology has been followed in the platform to obtain the data that will populate the entries. This methodology, which works with reasonable efficiency in the other languages – although it requires a supervised phase of correction and validation – entails an additional effort in the case of the Chinese language where, for example, discrepancies in the assignment of grammatical categories may affect the efficiency of the method in extracting candidates.
Keywords: collocations; phraseography; electronic lexicography; Mandarin Chinese

Cómo citar: Pazos Bretaña, J. M.; Orenha Ottaiano, A.; Xiong, Z. (2023). PLATCOL, Plataforma Multilingüe de Diccionarios de Colocaciones: el caso del chino. *Estudios de Traducción*, 13, 73-85.

[1] A partir del proyecto "A Phraseographical Methodology and Model for an Online Corpus-Based Multilingual Collocations Dictionary Platform" (Proceso FAPESP 2020/01783-2).
[2] Universidad de Granada
E-mail: jmpazos@ugr.es
ORCID: https://orcid.org/0000-0002-2429-8484
[3] Universidade Estadual Paulista "Júlio de Mesquita Filho" (Brasil)
E-mail: adriane.ottaiano@unesp.br
ORCID: https://orcid.org/0000-0001-8417-5120
[4] Universidad de Granada
E-mail: meixiong@correo.ugr.es
ORCID: https://orcid.org/0000-0001-9624-8550

1. Introducción

En las dos últimas décadas, las colocaciones han ocupado un lugar destacado en la agenda de la enseñanza y el aprendizaje de lenguas extranjeras (véanse Nesselhauf 2005; Alonso-Ramos 2008, 2019; Laufer 2011, Orenha-Ottaiano 2021; Torner & Bernal 2017, entre otros). A pesar de este hecho, cuando se trata de la traducción de colocaciones, el número de estudios que pueden contribuir a una mejor comprensión de las dificultades relativas a la complejidad de la traducción de dichas combinaciones no es tan significativo (Kenny 2001; Bernardini 2007; de Gregorio-Godeo y Molina 2011; Orenha-Ottaiano 2009, 2012, 2021).

Además, aunque varios autores destacan la importancia de la elaboración de diccionarios con un enfoque especial en las colocaciones o para la construcción de diccionarios de colocaciones específicos (Alonso-Ramos 2001, Atkins y Rundell 2008, Moon 2008; Orenha-Ottaiano 2013, 2016, 2017; Kilgarriff 2015, etc.), el número de diccionarios de colocaciones en línea o electrónicos disponibles es todavía escaso, especialmente cuando se trata de diccionarios de colocaciones bilingües o multilingües para la lengua general.

Este trabajo pretende llenar ese vacío mediante la descripción de una metodología para el diseño y la compilación de una plataforma en línea de diccionarios de colocaciones multilingües (inglés, portugués, francés, español y chino), PLATCOL, y la explicación de los problemas específicos encontrados durante el proceso de integración de la lengua china en el diccionario. La recopilación de colocaciones relevantes, en todas las lenguas, se basa en corpus y es semiautomática (extracción automática con validación humana). Además, el diseño de la plataforma tiene en cuenta las necesidades de los usuarios, tal como sugieren los principios de la teoría funcional de la lexicografía (Bothma y Tarp 2012, Fuertes-Olivera y Tarp 2014, Tarp 2015).

2. PLATCOL. Motivación y objetivos

PLATCOL es un diccionario multilingüe de colocaciones compuesto por un conjunto de diccionarios monolingües de colocaciones, uno para cada lengua incorporada (portugués, español, inglés, francés y chino). Se presenta al usuario a través de una interfaz de consulta web. Esta interfaz permite realizar búsquedas en modalidades monolingüe o multilingüe, dado que las entradas en diferentes lenguas están vinculadas entre sí.

2.1. Colocación

La literatura sobre colocaciones muestra que existen dos enfoques muy distintos para llevar a cabo la identificación y definición de colocaciones. En nuestro proyecto, definimos "colocación" guiándonos por ambos enfoques.

Bajo un enfoque estadístico, consideramos las colocaciones como combinaciones de palabras frecuentes, en una lengua específica, cuya coocurrencia a cierta distancia entre sí es estadísticamente más alta de lo esperado en comparación con una combinación aleatoria (Barfield y Gyllstad 2009, Nesselhauf 2005; Sinclair 1966, 1991, etc.). Sin embargo, como menciona Teubert (2004: 188), ser estadísticamente significativo no es suficiente para identificar una combinación de palabras como colocación: "También tienen que ser semánticamente relevantes. Deben tener un significado propio, un significado que no es obvio a partir del significado de las partes que las componen[5]".

Por esta razón, es importante describir las colocaciones bajo un enfoque fraseológico, y definimos las colocaciones como combinaciones omnipresentes, recurrentes y convencionalizadas que consisten en una base y un colocativo (Haussmann 1979, 1989), y poseen cierta fijación y restricción léxica o sintáctica. Se puede decir que son parcialmente composicionales porque su base mantiene su significado, no obstante, el colocativo puede adquirir un significado especial solo en combinación con la base (Alonso-Ramos 1994, Corpas 1996, Haussmann 1989 Heylen y Maxwell 1994, Orenha-Ottaiano 2020, Pamies 2019, Penadés Martínez 2017).

Así, en PLATCOL utilizamos una aproximación heurística, aplicamos primero un filtro estadístico para ordenar las combinaciones por los valores de las medidas de asociación y, posteriormente, inspeccionamos las combinaciones candidatas para eliminar las combinaciones libres o aleatorias. Este proceso se lleva a cabo en cada una de las lenguas incluidas.

2.2. Motivación y objetivos

El desarrollo de la competencia colocacional es una de las tareas más difíciles tanto para los estudiantes de Lenguas extranjeras (LE) como para los enseñantes de LE en formación. Esta dificultad también puede afectar a profesores en activo, que siguen enfrentándose al mismo problema incluso después de haberse licenciado,

[5] "They also have to be semantically relevant. They have to have a meaning of their own, a meaning that isn't obvious from the meaning of the parts they are composed of".

según nuestra experiencia de 20 años como profesores universitarios tanto en cursos de grado (Lenguas y Traducción) como de posgrado.

Como principal factor de motivación para la creación de PLATCOL, señalamos que, entre las muchas formas de ayudar a los estudiantes de LE y a los profesores en activo y en formación a alcanzar una competencia colocacional, el uso de diccionarios de colocaciones en línea puede considerarse una importante herramienta pedagógica.

En general, las colocaciones se mencionan, directa o indirectamente, como objeto de estudio en diferentes investigaciones que abordan el tema de la lexicografía pedagógica (Higueras 2005, 2006; Pérez Serrano 2014, Torner y Bernal 2017), lo que confirma la relevancia de este tipo de unidades para la enseñanza y el aprendizaje de lenguas extranjeras. En cuanto a la labor lexicográfica, existen excelentes diccionarios de colocaciones para estudiantes de LE, sin embargo, todavía existen carencias en la disponibilidad o publicación de diccionarios en línea.

Un objetivo principal de este proyecto se refiere al desarrollo de una plataforma que ofrezca un mayor grado de personalización de la estructura de los diccionarios: posibilidad de realizar consultas monolingües o multilingües a través de las lenguas incorporadas. Otros objetivos son el desarrollo de una metodología lexicográfica innovadora y un modelo para un diccionario de colocaciones multilingüe, así como el diseño de un sistema informático y una plataforma de colocaciones, PLATCOL, y, la creación de un recurso útil y amplio para la recuperación semiautomática de colocaciones, así como la extracción automática de buenos ejemplos, definiciones y traducción.

3. Estructura y diseño del diccionario de colocaciones multilingüe

El diccionario de colocaciones multilingüe (PLATCOL) propuesto aquí tiene como objetivo satisfacer las necesidades de los usuarios en cuanto a la codificación de la lengua y, como tal, se considera diccionario de producción. Además de ayudar a los usuarios a producir textos más idiomáticos, PLATCOL también tiene el propósito de desarrollar la competencia colocacional, que está intrínsecamente relacionada con la idiomaticidad. Una mayor competencia colocacional permitirá al alumno expresarse de manera más idiomática. Además, el diseño de PLATCOL también permite mostrar diccionarios monolingües. Así, servirá como diccionario monolingüe, bilingüe o multilingüe (inglés, portugués, francés, español y chino), dado que las entradas en diferentes lenguas están vinculadas entre sí.

PLATCOL está concebido como un diccionario en línea, el proyecto se inició con un diseño preliminar para servir de campo de pruebas. Una nueva versión ya está en una fase avanzada de construcción, se adaptará a las lenguas de incorporación más reciente (francés, español y chino), con un diseño más ambicioso e interactivo, así como unas características lexicográficas y una metodología más detalladas y mejoradas. El acceso está todavía restringido a los miembros del equipo, aunque la intención es abrirlo al público tan pronto como esté finalizado.

3.1. Perfil y necesidades del usuario

Frecuentemente, en los trabajos lexicográficos se hace referencia a los siguientes temas: tipología de los usuarios, sus necesidades y competencias. Así, en muchos estudios, el "problema" de los usuarios y sus necesidades son el foco principal. Sin embargo, como afirman claramente Fuertes Olivera y Tarp (2014), esta preocupación no da sus frutos, ya que no se materializa en decisiones teóricas y prácticas concretas, sino que los investigadores tienden a abordar el problema de forma más general y no profundizan en su discusión.

Nuestro trabajo se inscribe en la teoría funcional de la lexicografía y, por tanto, los siguientes constituyen puntos esenciales que guían el desarrollo de la plataforma: a) La definición previa de los perfiles de los usuarios a los que se dirige la propuesta, un paso crucial antes de su elaboración. Se trata de los perfiles ya definidos en Orenha-Ottaiano et al. (2021: 12) y que se detallan en la Tabla 2:

Tabla 2. Perfiles de usuario

Aprendientes de idiomas	Usuarios no nativos, estudiantes de una lengua extranjera de nivel intermedio o avanzado (a partir del nivel B1 en adelante, según el Marco Común Europeo de Referencia para las Lenguas: Aprendizaje, enseñanza, evaluación), en cualquier entorno (estudios universitarios, cursos de idiomas etc.)
Profesores en formación	Estudiantes en centros de enseñanza superior para convertirse en profesores de idiomas profesionales
Profesores	Profesores de idiomas como lengua extranjera, nativos o no, que se han formado para la enseñanza de LE

Estudiantes de traducción	Estudiantes de traducción en centros de enseñanza superior
Traductores	Traductores profesionales, nativos o no, de textos no especializados
Desarrolladores de materiales	Autores de manuales y materiales didácticos destinados a la enseñanza y el aprendizaje de lenguas
Lexicógrafos o investigadores	Investigadores en general, especialmente lingüistas, fraseólogos y lexicógrafos

Y b): La consideración de situaciones específicas extralexicográficas o sociales que motivarían el uso de la plataforma: "determinar qué tipo de necesidades puede tener un tipo concreto de usuario en cada tipo de situación" (Bergenholtz y Tarp 2003: 173).

Partimos de la idea de que los diferentes destinatarios de una obra lexicográfica tienen una serie de necesidades de información y consulta (Fuertes Olivera y Tarp 2014). Estas necesidades solo pueden satisfacerse si los usuarios tienen un acceso rápido y sencillo a un conjunto de datos lexicográficos elaborados según su perfil. De este modo, los usuarios deben poder extraer la información que necesitan, para poder emplearla posteriormente, según sus propósitos. Estos propósitos, a su vez, están siempre relacionados con los contextos y situaciones extra-lexicográficas que dieron lugar a estas necesidades (Tarp 2015).

Considerando el perfil de los potenciales usuarios de la plataforma, reconocemos que las situaciones sociales lexicográficamente relevantes, entre las cuatro definidas dentro de la teoría funcional, son las siguientes: 1. Comunicativas, en las que los usuarios tratan de resolver problemas relacionados con la producción, recepción, traducción, revisión y corrección de textos escritos u orales; y 2. Cognitiva, cuando los usuarios necesitan o quieren ampliar sus conocimientos sobre algo. Esta tipología podría aplicarse al perfil de todos los usuarios indicados; sin embargo, reconociendo las limitaciones de la propuesta, es necesario establecer algunas restricciones, como muestra la Tabla 3.

Tabla 3. Perfiles de usuario relacionados con situaciones sociales lexicográficamente relevantes y algunas restricciones (Orenha-Ottaiano et al. 2021:12)

Estudiantes de idiomas	Las situaciones comunicativas se limitan a la producción de textos escritos. En cuanto a las situaciones cognitivas, estarían relacionadas con el contexto de aprendizaje de la lengua
Profesores en formación	En el caso de los profesores no nativos en formación, las situaciones comunicativas están relacionadas con producción, traducción y corrección de textos escritos u orales. En cuanto a las situaciones cognitivas, nuestro objetivo es que los profesores en formación utilicen la plataforma para desarrollar la competencia colocacional
Profesores	En este caso, las situaciones comunicativas están relacionadas con la corrección y revisión de textos, y el diseño de actividades de desarrollo de la competencia colocacional. En el caso de los profesores no nativos, también pueden darse situaciones cognitivas, principalmente relacionadas con la preparación de materiales didácticos.
Traductores	En este caso, la Plataforma podría ser útil en muchas situaciones comunicativas –tanto en la recepción y en la traducción y revisión de textos como en situaciones cognitivas, para ayudar a los traductores que necesiten datos lexicográficos específicos relacionados con la frecuencia o el contexto de uso de una colocación, por ejemplo.
Desarrolladores de materiales	Las situaciones comunicativas relevantes para estos usuarios se refieren, sobre todo, a la obtención de información para la producción de materiales. También, en este caso, pueden darse situaciones cognitivas relacionadas con la elaboración de manuales y materiales de enseñanza-aprendizaje.
Lexicógrafos o investigadores	En el caso de los investigadores no nativos, las situaciones comunicativas pueden darse en situaciones relacionadas con la producción, revisión y corrección de textos. Lexicógrafos o investigadores nativos, por su parte, pueden encontrarse en contextos en los que la plataforma puede ser útil para acceder a cierta información sobre colocaciones como ejemplos, contextos de uso, clasificación, etc.

3.2. Macroestructura de los diccionarios

El conjunto de entradas, o nomenclatura, de PLATCOL está formado, para cada lengua, por listas de palabras clave que pertenecen a tres clases morfosintácticas: sustantivos, verbos y adjetivos (véase sección 4.2). Incluye, además, una introducción sistemática y una guía de uso.

3.3. Microestructura de los diccionarios

La compilación de un diccionario de colocaciones, una tarea ya de por sí compleja, se convierte en un reto aún mayor cuando se tienen en cuenta varias lenguas. La organización de la microestructura, como se explica a continuación, es especialmente laboriosa

Las entradas de PLATCOL incluyen sustantivos, verbos y adjetivos que corresponden a las bases, o colocativos, de las colocaciones (véase más sobre las estructuras de las colocaciones en esta sección).

En un diccionario de colocaciones, las palabras principales pueden organizarse según al menos dos principios diferentes. Uno de los puntos de vista en el tratamiento de las colocaciones se basa en la estadística. Las colocaciones se definen bajo un enfoque estadístico en función de su frecuencia de coocurrencia. De este modo, la palabra clave puede ser la base o el colocativo.

El otro punto de vista sigue el enfoque de Hausmann (1985, 1989), que utiliza el concepto de base, el elemento habitualmente conocido por los usuarios, y de colocativo, el elemento que buscan, es decir, lo que necesitan encontrar los estudiantes y los traductores, por ejemplo.

En este proyecto, hemos optado por esta última visión (Hausmann 1985, 1989), alegando que es más fácil de usar y eficaz, considerando la mayoría de los perfiles de usuario, además de ser el punto de partida para muchos de los usuarios. Por otra parte, los usuarios podrán realizar búsquedas de bases o de colocativos en la barra de búsqueda de la plataforma. Las entradas de los diccionarios de colocaciones multilingües constan de los siguientes elementos:

Tabla 4. Elementos de estructuración de las entradas (Orenha-Ottaiano et al. 2021:13)

Una **entrada,** que corresponde a la base de las colocaciones. Las entradas pueden ser sustantivos, verbos o adjetivos.
Clase de palabra: la clase de la palabra se coloca justo después de la entrada. En el caso de estos diccionarios de colocaciones, serán un sustantivo (n.), un verbo (v.) o un adjetivo (adj.). Si una palabra pertenece a más de una clase de palabras, como, por ejemplo, *abstract* (n.), *abstract* (v.) y *abstract* (adj.), en inglés, cada clase de palabra aparece en entradas separadas, de modo que las colocaciones, estructuras de colocaciones y otras informaciones se organizan fácilmente.
Frecuencia de cada entrada.
Definición: se ofrecerá una breve definición de los diferentes sentidos de cada base. La decisión de incluir una definición es que las colocaciones puedan ser debidamente organizadas de acuerdo con cada sentido de la entrada. Por lo tanto, los usuarios podrán tener un acceso más rápido a las colocaciones que están buscando.

Tabla 5. Organización de las colocaciones (Orenha-Ottaiano et al. 2021:14)

Estructura sintáctica	dependiendo de la parte de la oración de la entrada y de la lengua de la colocación, se organizan según la estructura sintáctica (Hausmann 1985, 1989, Orenha-Ottaiano 2009, 2016, 2017)
Taxonomía	verbales, nominales, adjetivas y adverbiales.
Visualización de las entradas	En cada sección de cada palabra clave y definición, los usuarios pueden elegir que las colocaciones se muestren en orden alfabético o por frecuencia o saliencia (clasificadas según su puntuación estadística). Para usuarios más especializados, como investigadores y lexicógrafos, las colocaciones del diccionario pueden clasificarse por los valores de las medidas de asociación estadística.
Incorporación de ejemplos de uso	Para ilustrar cómo se utilizan las colocaciones en función de un significado concreto, los usuarios tendrán la posibilidad de elegir la visualización de hasta 5 ejemplos.

4. Metodología

La metodología para construir cada diccionario monolingüe se basa en el enfoque automático descrito en García et al. (2019a), enriquecido con información de sentido de las bases y una revisión y validación manual de los datos extraídos que es realizada por lexicógrafos.

4.1. Corpus

Recopilamos un gran corpus para cada una de las cinco lenguas del proyecto utilizando diferentes datos de origen, como muestra el cuadro siguiente:

	Fuentes
Portugués	Jornal do Brasil, Wikipedia/Wikibooks, Paracrawl, CHAVE (Santos y Rocha 2005), CBras, BrWaC (Wagner Filho et al. 2018)
Español	EuroParl (Koehn 2005), Literatura (relatos breves/romántica) (García et al. 2019a), Wikipedia/Wikibooks
Inglés	EuroParl, Wikipedia/Wikibooks
Francés	FrWaC (Baroni et al. 2009), Wikipedia/Wikibooks
Chino	Literatura, Wikipedia/Wikibooks

Los corpus se analizaron con UDPipe (Straka y Straková 2017) utilizando los últimos modelos (v2.7) entrenados en los corpus UD (de Marneffe et al. 2021). Previo a este análisis sintáctico, lematizamos y etiquetamos (PoS) los datos utilizando los mismos modelos de UDPipe para el inglés y el francés, LinguaKit (Gamallo et al. 2018) para el portugués y el español, y la suite Stanford CoreNLP (Manning et al. 2014) para los textos chinos.

4.2. Definición y extracción de palabras clave

Nos centramos en los tipos de colocaciones con tres clases morfosintácticas de bases: sustantivos, verbos y adjetivos. Debido al gran tamaño de los corpus, intentamos extraer listas de palabras clave para cada clase e idioma. Por lo tanto, procedimos a la extracción automática de los lemas con una frecuencia mínima de una ocurrencia por millón de *tokens* en cada corpus, anotándolos como conocidos o desconocidos si aparecen en léxicos grandes. Utilizamos los diccionarios proporcionados por FreeLing (Padró y Stanilovsky 2012) para cada idioma (inglés, portugués, francés y español), excepto para el chino, puesto que no conocemos ningún diccionario gratuito para esta lengua.

Tras la extracción automática, que se realizó para cada idioma por separado, las listas de palabras clave se sometieron a los lexicógrafos para filtrar el ruido (por ejemplo, lemas con erratas, entradas mal procesadas, etc.) y para seleccionar los lemas más frecuentes, que luego se utilizaron para extraer las colocaciones candidatas. Además, cada palabra clave se enriqueció semánticamente con los posibles sentidos presentes en WordNet, utilizando la Open Multilingual WordNet (Bond y Foster 2013) mediante la interfaz proporcionada por el paquete NLTK (Bird y Klein 2009).

4.3. Identificación de colocaciones y oraciones de ejemplo

Siguiendo a Garcia et al. (2017), extrajimos los pares de las relaciones de dependencia objetivo utilizando las palabras clave validadas manualmente y restringiendo las posibles colocaciones por su categoría morfosintáctica. Así, para las bases sustantivas extrajimos las siguientes relaciones sintácticas[6]: *obj* (colocaciones verbo-sustantivo), *nsubj* (instancias de sustantivo-verbo), *obl* (verbo-preposición-sustantivo), *amod* (adjetivo-sustantivo), y *nmod* y *compound* (ambas incluyen instancias de sustantivo-sustantivo o sustantivo-prep-sustantivo). Para las bases verbales extrajimos *xcomp* (colocaciones verbo-adjetivo) y *advmod* (verbo-adverbio). Por último, para las bases adjetivas, extrajimos ejemplos *advmod* (adjetivo-adverbio).

Para cada triplete (base; colocativo; relación: véanse las relaciones sintácticas descritas en el párrafo anterior) seguimos el método de coocurrencia sintáctica descrito en Evert (2008) para calcular, además de los datos de frecuencia, los siguientes valores estadísticos: PMI, Dice, log-likelihood, t-score, z-score, 2 y simple-ll, junto con ΔP (Gries 2013). Para reducir el gran tamaño de los conjuntos de candidatos, eliminamos aquellas combinaciones con una frecuencia normalizada inferior a uno por millón, y ordenamos las restantes por t-score (Garcia et al. 2019b).

A continuación, recogimos hasta ocho frases para cada colocación candidata, seleccionadas por un conjunto de heurística inspirada en GDEX (Kilgarriff et al. 2008). Implementamos una estrategia básica utilizando algunas de las propuestas de Kosem et al. (2019) para el inglés y para el portugués (estas últimas también se utilizaron para las demás lenguas romances): las oraciones con menos de seis *tokens* se descartaron y las que tienen más de 30 *tokens* se penalizaron de forma incremental. Además, también se penalizaron las frases con puntuación, nombres propios, palabras con más de 12 caracteres y caracteres extraños (por ejemplo, en otros alfabetos y codificaciones). No se han aplicado otras heurísticas de la bibliografía porque requieren recursos específicos de la lengua o son muy costosas desde el punto de vista computacional.

Esta información extraída automáticamente fue utilizada por expertos lingüistas para seleccionar las colocaciones para el recurso final. Para cada candidato, los lexicógrafos deciden qué combinaciones se van

[6] https://universaldependencies.org/u/dep/

a incorporar al diccionario, el criterio principal es la eliminación de combinaciones libres, y seleccionan el sentido adecuado para la base y un conjunto de cinco ejemplos que se mostrarán en la plataforma.

El volumen de los datos recuperados automáticamente es muy grande. Hemos establecido un filtro de 20 ocurrencias por millón, en la misma dependencia sintáctica, siguiendo a Evert (2008). Este filtro ha proporcionado, de media, 20.000 candidatos con base = nombre, y 8.000 con base = verbo, por ejemplo. La fase de post edición está todavía en curso y puede durar algunos meses, ya que los datos se validan, evalúan y revisan manualmente por al menos dos lexicógrafos. A medida que las colocaciones son revisadas, se dirigen a la siguiente fase de asociación automática con equivalencias en otras lenguas, como se describe en la siguiente sección, según los pares que hemos establecido previamente (véase la subsección 4.2.).

4.4. Propuesta de equivalencias en otras lenguas

Una vez incluidas las colocaciones correspondientes a cada palabra clave en los diccionarios monolingües de la plataforma, utilizaremos un enfoque no supervisado para recuperar las equivalencias candidatas entre las lenguas del proyecto. La estrategia, inspirada en Garcia et al. (2019c), puede resumirse como sigue: primero entrenamos modelos monolingües word2vec (Mikolov et al. 2013) utilizando corpus procesados y representando cada palabra como un par de lema y etiqueta PoS (por ejemplo, "house_NOUN"). A continuación, estos modelos se mapean en un espacio vectorial compartido con vecmap (Artetxe et al. 2018). Por último, creamos un vector de composición para una determinada colocación en la lengua A, y buscamos candidatos similares (en términos de similitud del coseno) en la lengua B (Garcia et al. 2019c). Las traducciones candidatas se clasifican según la confianza de los modelos, y serán validadas manualmente por lexicógrafos en trabajos posteriores.

5. El chino en PLATCOL

La integración del chino estándar (mandarín) en PLATCOL comenzó en 2020 con la selección y creación de un corpus que serviría para extraer las palabras clave y candidatos a colocación. La organización, interfaz y posibilidades de consulta es idéntica para todas las lenguas incluidas. El corpus chino consta de aproximadamente 600 millones de *tokens*. Está formado por dos grupos de textos

a. 598 libros en formato electrónico, en ficheros de texto con caracteres simplificados y codificación *unicode*, que pertenecen a géneros de ficción y no-ficción.
b. El texto de la versión china de *Wikipedia*[7].

El método para obtener los candidatos sigue, en general, los principios metodológicos descritos en la sección 4. Sin embargo, al aplicarlo a la lengua china, nos enfrentamos a desafíos significativos que requieren una adaptación y consideración cuidadosa. Por lo tanto, llevamos a cabo una serie de experimentos que nos permitieron evaluar la eficacia de la metodología y abordar los desafíos específicos que surgen al aplicarla a la lengua china.

5.1. Experimentos

5.1.1. Primer experimento: Implementación de la metodología general

En el primer experimento, procedimos a implementar directamente la metodología general para identificar colocaciones en chino. En el resultado se observó bastante ruido, evidenciándose la presencia de una serie de combinaciones erróneas, como, por ejemplo, *shǒu dū* + **shè jù* 首都 + *设具 ('capital + *herramienta'), *jiāo gěi* + * *bá yù* 交给+跋郁 ('entregar + **bayu*'), entre otras. Presumimos que esto podría atribuirse a la segmentación inexacta de las palabras, dado que el sistema de escritura chino no utiliza espacios gráficos que funcionen como separadores entre secuencias de caracteres para delimitar palabras. Con el fin de abordar esta problemática, llevamos a cabo un segundo experimento con el objetivo de encontrar una solución más eficiente.

5.1.2. Segundo experimento: Cambio de herramienta de segmentación

En el segundo experimento, decidimos cambiar la herramienta de segmentación utilizada, pasando de UDPipe (que ha sido utilizada para otras lenguas en PLATCOL) a Stanford CoreNLP, lo cual implicó llevar a cabo una nueva *tokenización* del texto. En este caso, observamos, tras una revisión realizada por hablantes nativos,

[7] https://zh.wikipedia.org/

que la calidad de las combinaciones extraídas mejoró de manera sustancial. No obstante, se observaron demasiadas combinaciones problemáticas entre las combinaciones candidatas a colocación extraídas. Es decir, se identificaron tres categorías de problemas específicos que requieren atención detallada:

a. el problema principal fue debido a una etiquetación incorrecta: p.ej., en la tabla de "colocación con adjetivo" los componentes de la colocación deben incluir un adjetivo en lugar de otra categoría. Por ejemplo: *mò rán* + *huí shǒu* 蓦然+ 回首 ('de repente + mirar atrás');
b. el segundo desafío fue causado por la extracción de una serie de combinaciones libres: p.ej.: *bù* + *dòng shǒu* 不+动手 ('no + hacer'), que es un sintagma libre producto de una regla gramatical, de modo que no debería ser incluido;
c. el tercer problema fueron interferencias debidas a locuciones, ejemplos erróneos como *xǐ xīn yàn jiù* 喜新厌旧 (abandonar lo viejo por lo nuevo), que aparece en la lista de adjetivos, cuando se trata de una locución verbal en lugar de colocación adjetival.

5.1.3. Tercer experimento: Consideraciones léxico-semánticas

En respuesta a los desafíos mencionados, realizamos un tercer experimento para abordar las particularidades lingüísticas de la lengua china, teniendo en cuenta la flexibilidad en la categoría gramatical y la incorporación de criterios léxico-semánticos para la selección de colocaciones correctas.

Para empezar, conviene señalar que la lengua china cuenta con pocas inflexiones gramaticales, lo que resulta en cambios de categoría gramatical sin modificaciones morfológicas, fenómeno conocido como "palabra multifuncional" o "conversión" en la comunidad lingüística china (Qiao 2017).

Con objeto de reducir la ambigüedad generada por esta peculiaridad, procedimos a eliminar primero las bases etiquetadas incorrectamente. Para ello, realizamos una revisión manual, por parte de hablantes nativos chinos. Por ejemplo, la palabra *jīng shén* 精神 puede funcionar tanto como sustantivo "energía/vigor" como adjetivo "energético/vigoroso".

En lo concerniente a los problemas relacionados con la aparición de combinaciones libres y locuciones, optamos por aplicar un filtro manual basado en criterios léxico-semánticos con el fin de excluir aquellas expresiones que no sean colocaciones. Este proceso está siendo llevado a cabo por expertos chinos, siguiendo los siguientes criterios. Por una parte, las colocaciones adecuadas deben cumplir con alguna de las siguientes normas:

a. Combinaciones imprevisibles, es decir, combinaciones a las que los usuarios podrían desear acceder, por ejemplo, cuando se desea expresar la idea de que el talento se acaba, el verbo preferido en chino sería *kú jié* 枯竭 ('secarse'), es decir, con colocaciónes como *cái sī kú jié* 才思枯竭 (lit. ,'secarse el talento'; fig. 'quedarse sin talento o ideas');
b. Combinaciones que plantean dificultades significativas a la hora de producirse en el habla o la escritura, por ejemplo, a un aprendiente de chino como LE que quiera expresar la idea de "intensidad de interés", puede parecerle obvia la asociación con el adjetivo *qiáng liè* 强烈 ('gran, fuerte'), sin embargo, la asociación idiomática con *nóng hòu* 浓厚 ('espeso, grueso'), no es una opción evidente. Por ello, debe llamarse la atención sobre la colocación habitual *xìng qù nóng hòu* 兴趣浓厚 ('interés vivo'), para que pueda ser aprendida.
c. Combinaciones difíciles de traducir a otras lenguas;
d. Combinaciones léxicas restringidas con significado literal. Por ejemplo, no es apropiado reemplazar la expresión nóng chá 浓茶 ('té fuerte') por *qiáng chá* 强茶 ('té fuerte'). Aunque el significado de toda la combinación sea compositivo, debe incluirse en el diccionario (Torner & Bernal, 2017 167);

Por otra parte, las combinaciones que no deben seleccionarse son las siguientes:

a. Nombres propios (por ejemplo, nombres de ciudad, país, etc.);
b. Combinaciones muy especializadas, por ejemplo, *xiàng liàng kōng jiān* 向量空间 ('espacio vectorial');
c. Combinaciones libres o locuciones.

Solo aquellas colocaciones que cumplen con estos criterios están siendo seleccionadas y son consideradas válidas.

5.2. Análisis del resultado

5.2.1. Resultado

En el último experimento obtuvimos un listado de candidatos a colocaciones. Con el fin de evaluar la eficacia de la metodología, una de las métricas fundamentales es el índice de precisión. Específicamente, el índice de precisión se calcula dividiendo el número de colocaciones "correctas" entre el número total de candidatos. Sin embargo, debido al tamaño del corpus, resulta impracticable revisar todas las colocaciones. La práctica común consiste en seleccionar alguna palabra clave al azar y evaluar la eficacia de la base de datos examinando todos los colocativos de esa palabra.

Por ejemplo, en las investigaciones previas como las de Sun et al. (1997), Chen (2006), Qian (2012) y Zeng (2015) seleccionaron la palabra *néng lì* 能力 ('capacidad, habilidad') para evaluar el porcentaje de acierto de sus experimentos. Con el fin de posibilitar la comparación, en este experimento utilizamos también esta misma palabra.

En el listado de candidatos, la palabra *néng lì* (能力) aparece un total de 21.033 veces en nuestro corpus. Tras una verificación manual por hablantes nativos, observamos que 20.391[8] forman colocaciones correctas (un acierto del 96,95 %), distribuidas en diversas tipologías sintácticas. Entre ellas, 2.374 son del tipo *amod*, 3.116 de *nmod*, 1.028 del *nsubj*, 13.723 de *obj* y 150 del *obl*. Véase la distribución de los candidatos correctos en la tabla 6.

Tabla 6. Distribución de resultados de los candidatos sobre la base 能力 *néng lì* ('capacidad')

Tipología de candidatos (Relación de dependencia)	Frecuencia de aparición	Porcentaje
amod	2374	11.29 %
nmod	3116	14.81 %
nsubj	1028	4.89 %
obj	13723	65.25 %
obl	150	0.71 %
combinaciones falsas	642	3.05 %
Total	**21033**	**100.00 %**

A continuación, se presentan algunos ejemplos representativos de colocaciones correctas en la tabla 7, que incluyen información sobre la palabra clave y su colocativo, la relación de dependencia (la columna *deprel*), la frecuencia de la palabra clave en esa relación (*freqbase*), la frecuencia del colocativo (*freqcolocativo*), la frecuencia del candidato a colocación (*freq*) y su frecuencia normalizada (*freqnorm*). En la plataforma los usuarios podrán consultar esta información de acuerdo con sus necesidades.

Tabla 7. Muestra de colocaciones correctas sobre la base 能力 *néng lì* ('capacidad')

base	colocativo	deprel	freqbase	freqcolocativo	freq	freqnorm
能力	出色 'excepcional'	amod	6398	4974	154	45.61
能力	非凡 'extraordinaria'	amod	6398	868	61	18.07
能力	神奇 'mágica'	amod	6398	2436	61	18.07
能力	失去 'perder'	obj	29295	25000	829	57.74
能力	恢复 'recuperar'	obj	29295	15022	159	11.08
能力	提高 'aumentar'	obj	29295	14793	679	47.3

Al mismo tiempo, se presentan también unos ejemplos que corresponden a combinaciones que no cumplen con la noción de colocación en nuestra plataforma. Un caso es el extraído de la palabra *jìn xíng* 进行 'ejecutar', la cual forma una combinación poco idiomática, no muy común y que no está registrada en diccionarios de

[8] Estas colocaciones son instancias verificadas de combinaciones *palabra clave* + *palabra* (base + colocativo). Pueden ser consultadas en PLATCOL realizando una búsqueda por *palabra clave*. Para cada colocación *palabra clave* + *palabra* se proporcionarán ejemplos de uso.

colocaciones como el *Diccionario de Colocación del Chino Moderno* (*Xian dai han yu da pei ci dian* 现代汉语搭配词典) (Mei, 1999) con la base *néng lì* (能力, 'capacidad').

Tabla 8. Errores observados en la cata realizada sobre la base 能力 *néng lì* ('capacidad')

base	colocativo	deprel	freqbase	freqcolocativo	freq	freqnorm
能力	无 'no'	obj	29295	54791	277	19.29
能力	使 'dejar'	nsubj	8828	56650	166	16.84
能力	具 'proveer'	obj	29295	20829	145	10.1
能力	进行 'ejecutar'	obl	1465	37087	27	10.49
能力	是 'sí'	obl	1465	38234	27	10.49

5.2.2. Análisis comparativo

A continuación, comparamos los resultados con investigaciones previas. Sun et al. (1997) obtienen un total de 498 candidatos, de los cuales 169 se validaron como colocaciones, un índice de acierto del 33,94 %. Por su parte, Chen (2006) extrajo 275 candidatos con la palabra *néng lì*, de los que 233 eran colocaciones, obteniendo un índice de acierto del 84,73 %. Otro autor Qian (2012) llevó a cabo un estudio en el que se identificaron 153 candidatos a colocaciones verbales de las que se validaron 125 (81,70 %). Por último, Zeng (2015) obtuvo un total de 796 candidatos de los que 662 eran colocaciones, logrando así un 83,19%. Véase la siguiente tabla 9 de comparación.

Tabla 9. Comparación con otros experimentos sobre la base 能力 *néng lì* ('capacidad')

	Candidatos totales extraídos	Candidatos correctos	Porcentaje de acierto
Sun Maosong et al., 1997	498	169	33.94 %
Chen Yaju, 2006	275	233	84.73 %
Qian Xiaofei, 2012	153	125	81.70 %
Zeng Tong, 2015	796	662	83.19 %
Nuestro, 2022	21033	20391	96.95 %

En términos de metodología, las investigaciones anteriores se basaron principalmente en enfoques estadísticos. Por ejemplo, Zeng (2015) aclara que optó por descartar las consideraciones lingüísticas (p.ej., la información de categoría gramatical de las palabras, la distinción entre combinaciones libres y locuciones, etc.) y controló el procedimiento a través de los indicadores estadísticos de frecuencia de coocurrencia.

La diferencia en los resultados obtenidos puede atribuirse tanto al diseño del sistema y la metodología propios como a la incorporación de criterios lingüísticos, lo cual nos permite descartar candidatos no apropiados, mejorando así significativamente el índice de acierto en la identificación de colocaciones en la lengua china.

6. Conclusiones

A modo de conclusión, podemos afirmar que la metodología general utilizada en PLATCOL para la obtención de candidatos es válida también para la lengua china, con algunas adaptaciones en función de las particularidades de esta lengua. Por una parte, desde una perspectiva técnica, el análisis estadístico ha sugerido que el empleo de Standford CoreNLP resulta más preciso en comparación con UDPipe. Por otra parte, los datos obtenidos nos muestran que la calidad de los resultados es alentadora, pese a las labores adicionales requeridas debido a la naturaleza de la lengua china. Por último, conviene indicar que, para futuros trabajos, estimamos que es posible que mejore la calidad de los candidatos en chino con la incorporación de un *stopword list*.

Referencias

Alonso-Ramos, Margarita (1994). Hacía una definición del concepto de colocación: de J. R. Firth a I. A Mel'čuk. *Revista de Lexicografía, 1*, 9-28.

Alonso-Ramos, Margarita (2001). Construction d'une base de données des collocations bilingue français-espagnol. *Langages*, *35* (143), 5-27. https://doi.org/10.3406/lgge.2001.888

Alonso-Ramos, Margarita (2008). Papel de los diccionarios de colocaciones en la enseñanza de español como L2. En E. Bernal y J. DeCesaris (Eds.), *Proceedings of the XIII EURALEX International Congress* (pp. 1215-1230). IULA/Documenta Universitaria.

Alonso-Ramos, Margarita y García-Salido, Marcos (2019). Testing the Use of a Collocation Retrieval Tool Without Prior Training by Learners of Spanish. *International Journal of Lexicography*, *32* (4), 480-497. https://doi.org/10.1093/ijl/ecz016

Artetxe, Mikel, Labaka, Gorka y Agirre, Eneko (2018). A robust self-learning method for fully unsupervised crosslingual mappings of word embeddings. En Iryna Gurevych y Yusuke Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 1*, 789-798. https://doi.org/10.18653/v1/P18-1073

Atkins, B. T. Sue y Rundell, Michael (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

Barfield, Andy y Gyllstad, Henrik (Eds.) (2009). *Researching Collocations in another language: Multiple Interpretations*. Palgrave Macmillan.

Baroni, Marco; Bernardini, Silvia; Ferraresi, Adriano y Zanchetta, Eros (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43* (3), 209-226. https://doi.org/10.1007/s10579-009-9081-4

Bergenholtz, Henning y Tarp, Sven (2003). Two opposing theories: On H.E. Wiegand's recent discovery of lexicographic functions. *HERMES - Journal of Language and Communication in Business*, *31*, 171-196. https://doi.org/10.7146/hjlcb.v16i31.25743

Bernardini, Silvia (2007). Collocations in Translated Language: Combining Parallel, Comparable and Reference Corpora. En Matthew Davies, Paul Rayson, Susan Hunston y Pernilla Danielsson (Eds.), *Proceedings of the Corpus Linguistics Conference (CL2007)* (pp. 1-16). University of Birmingham. Disponible en: http://ucrel.lancs.ac.uk/publications/CL2007/paper/15_Paper.pdf.

Bird, Steven; Klein, Ewin y Loper, Edward (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Bond, Francis y Foster, Ryan (2013). Linking and extending an open multilingual Wordnet. En Hinrich Schuetze, Pascale Fung y Massimo Poesio (Eds.), *Proceedings of the 51st annual meeting of the association for computational linguistics (Volume 1: Long papers)* (pp. 1352-1362). Association for Computational Linguistics. Disponible en: https://www.aclweb.org/anthology/P13-1133.pdf

Bothma, Theo. J. D., y Tarp, Sven (2012). Lexicography and the Relevance Criterion. *Lexikos*, *22*, 86-108. https://doi.org/10.5788/22-1-999

Chen, Yaju. (2006). *Xian dai han yu ci yu da pei de zi dong chou qu fang fa* 现代汉语词语搭配的自动抽取方法 *[Método de extracción automática de colocaciones de palabras en chino moderno]*. East China Normal University.

Corpas Pastor, Gloria (1996). *Manual de fraseología española*. Gredos.

de Gregorio-Godeo, Eduardo y Molina, Silvia (2011). Collocations and the Translation of News: An English–Spanish Electronic Dictionary of Multi-Word Combinations as a Translation Tool. *Perspectives*, *19* (2), 135-152.

de Marneffe, Marie Catherine; Manning, Christopher D.; Nivre, Joakim y Zeman, Daniel (2021). Universal Dependencies. *Computational Linguistics*, *47* (2), 255-308. https://doi.org/10.1162/coli_a_00402

Evert, Stefan (2008). Corpora and collocations. En A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics* (Vol. 2, pp. 1212-1248). Mouton de Gruyter.

Filho Wagner, Jorge A., Wilkens, Rodrigo; Idiart, Marco y Villavicencio, Aline (2018). The brWaC Corpus: A New Open Resource for Brazilian Portuguese. En K. C. Nicoletta, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis y T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 4339-4344). European Language Resources Association. Disponible en: https://aclanthology.org/L18-1686

Fuertes-Olivera, Pedro Antonio y Tarp, Sven (2014). *Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminography*. De Gruyter. https://doi.org/10.1515/9783110349023

Gamallo, Pablo, Garcia, Marcos, Piñeiro, César, Martínez-Castaño, Rodrigo y Pichel, Juan C. (2018). LinguaKit: A big data-based multilingual tool for linguistic analysis and information extraction. *Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 239-244. https://doi.org/10.1109/SNAMS.2018.8554689

Garcia, Marcos; García-Salido, Marcos y Alonso-Ramos, Margarita (2017). Using bilingual word-embeddings for multilingual collocation extraction. En S. Markantonatou, C. Ramisch, A. Savary y V. Vincze (Eds.), *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)* (pp. 21-30). ACL. https://doi.org/10.18653/v1/W17-1703

Garcia, Marcos; García-Salido, Marcos y Alonso-Ramos, Margarita (2019a). A comparison of statistical association measures for identifying dependency-based collocations in various languages. En A. Savary, C. P. E. Agata, F. Bond, J. Mitro-vić y V. B. Mititelu (Eds.), *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)* (pp. 49-59). ACL. https://doi.org/10.18653/v1/W19-5107

Garcia, Marcos; García-Salido, Marcos y Alonso-Ramos, Margarita (2019b). Towards the automatic construction of a multilingual dictionary of collocations using distributional semantics. En I. Kozem, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, C. Tiberius y T. Zingano Kuhn (Eds.), *Proceedings of eLex 2019: Smart Lexicography* (pp. 747-762). Lexical Computing CZ. Disponible en: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_42.pdf

Garcia, Marcos; García-Salido, Marcos y Alonso-Ramos, Margarita (2019c). Weighted compositional vectors for translating collocations using monolingual corpora. En G. Corpas Pastor & R. Mitkov (Eds.), *Computational and Corpus-Based Phraseology* (pp. 113-128). Springer. https://doi.org/10.1007/978-3-030-30135-4_9

Gries, Stephan Th. (2013). *Statistics for linguistics with R: a practical introduction* (2nd revised). De Gruyter Mouton. https://doi.org/10.1515/9783110718256

Hausmann, Franz Josef (1985). Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. En H. Bergenholtz y J. Mugdan (Eds.), *Lexikographie und Grammatik* (pp. 118-129). De Gruyter. https://doi.org/10.1515/9783111635637-004

Hausmann, Franz Josef (1989). Le dictionnaire de collocations. En O. Reichmann, H. E. Wiegand y L. Zgusta (Eds.), *Wörterbücher: Ein internationales Handbuch zur Lexicographie. Dictionaries. Dictionnaires* (pp. 1010-1019). De Gruyter.

Heylen, Dirk y Maxwell, Kerry (1994). Lexical functions and the translation of collocations. *International Conference on Computational Linguistics*, Kyoto, Japan, pp. 298-305.

Higueras-García, Marta (2005). Necesidad de un diccionario de colocaciones para aprendientes de ELE. En M. A. Castillo et al. (Eds.). *Las gramáticas y los diccionarios en la enseñanza del español como segunda lengua: deseo y realidad*. Actas del XV Congreso Internacional de ASELE (pp. 480-490). Universidad de Sevilla.

Jousse, Anne-Laure y Polguère, Alain (2005). *Le DiCo et sa version DiCouébe. Document descriptif et manuel d'utilisation*. Université de Montréal: Observatoire de linguistique Sens-Texte (OLST). Disponible en: http://idefix.ling.umontreal.ca/dicouebe/DiCoDOC.pdf

Kenny, Dorothy (2001). *Lexis and creativity in translation: A corpus-based study*. St. Jerome Pub. https://doi.org/10.4324/9781315759968

Kilgarriff, Adam; Husák, Miloš; McAdam, Katy; Rundell, Michael y Rychly, Pavel (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. En E. Bernal y J. DeCesaris (Eds.), *Proceedings of the 13th EURALEX International Congress* (pp. 425-432). Institut Universitari de Linguistica Aplicada. Universitat Pompeu Fabra. Disponible en: https://euralex.org/publications/gdex-automatically-finding-good-dictionary-examples-in-a-corpus/

Kilgarriff, Adam; Marcowitz, Fredrik; Smith, Simon y Thomas, James (2015). Corpora and Language Learning with the Sketch Engine and SKELL. *Revue française de linguistique appliquée*, *XX* (1), 61-80. https://doi.org/10.3917/rfla.201.0061

Koehn, Philipp (2005). Europarl: A parallel corpus for Statistical Machine Translation. *Proceedings of the 10th Machine Translation Summit*, 79-86. Disponible en: https://aclanthology.org/2005.mtsummit-papers.11

Kosem, Iztok, Koppel, Kristina; Kuhn, Tanara Z.; Michelfeit, Jan y Tiberius, Carole (2019). Identification and automatic extraction of good dictionary examples: The case(s) of GDEX. *International Journal of Lexicography*, *32* (2), 119-137. https://doi.org/10.1093/ijl/ecy014

Laufer, Batia (2011). The Contribution of Dictionary Use to the Production and Retention of Collocations in a Second Language. *International Journal of Lexicography*, *24* (1), 29-49. https://doi.org/10.1093/ijl/ecq039

Manning, Christopher; Surdeanu, Mihai; Bauer, John; Finkel, Jenny; Bethard, Steven y McClosky, David (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55-60. https://doi.org/10.3115/v1/P14-5010

Mei, Jiaju (Ed.) (1999). *Xian dai han yu da pei ci dian* 现代汉语搭配词典 *['Diccionario de Colocación del Chino Moderno']. Shanghai: Han yu da ci dian chu ban she* 汉语大词典出版社.

Mikolov, Tomas; Chen, Kai; Corrado, Greg y Dean, Jeffrey (2013). Efficient estimation of word representations in vector space. En Y. Bengio y Y. LeCun (Eds.), *Workshop Proceedings of the International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1301.3781

Moon, Rosamund (2008). Sinclair, Phraseology, and Lexicography. *International Journal of Lexicography*, *21* (3), 243-254. https://doi.org/10.1093/ijl/ecn027

Nesselhauf, Nadja (2005). *Collocations in a Learner Corpus*. John Benjamins. https://doi.org/10.1075/scl.14

Orenha-Ottaiano, Adriane (2009). A compilação de corpora comparáveis na área de negócios e sua relevância para a tradução e terminologia. *Calidoscópio, 7* (3), 232-36.

Orenha-Ottaiano, Adriane (2012). English collocations extracted from a corpus of university learners and its contribution to a language teaching pedagogy. *Acta Scientiarum*, *34* (2), 241-251.

Orenha-Ottaiano, Adriane (2013). The proposal of an electronic bilingual dictionary based on corpora. En O. M. Karpova (Ed.), *Life Beyond Dictionaries. Proceedings of X Anniversary International School on Lexicography* (pp. 405-408).

Orenha-Ottaiano, Adriane (2016). The compilation of a printed and online corpus-based bilingual collocations dictionary. En G. Meladze, T. Margalitadze e I. Javakhishvili (Eds.)*, Proceedings of the 17th EURALEX international congress* (pp. 735-745). Tbilisi University Press.

Orenha-Ottaiano, Adriane (2017). The compilation of an Online Corpus-Based Bilingual Collocations Dictionary: motivations, obstacles and achievements. En I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek y V. Baisa (Eds.), *Proceedings of eLex 2017–Electronic lexicography in the 21st century: Lexicography from Scratch* (pp. 458-473). Lexical Computing CZ, s.r.o.

Orenha-Ottaiano, Adriane (2020). The creation of an online English collocations platform to help develop collocational competence. *Phrasis: Revista di studi fraseologici e paremiologic. Associazone Italiana di Fraseologia e Paremiologia, 1*, 59-81.

Orenha-Ottaiano, Adriane (2021). Escollas colocacionais a partir dun corpus de estudantes de tradución e a importancia do desenvolvemento da competencia colocacional. *Cadernos de Fraseoloxía Galega, 21*, 35-64.

Orenha-Ottaiano, Adriane; Garcia, Marcos; Olímpio De Oliveira, Maria Eugênia; L'Homme, Marie-Claude; Alonso Ramos, Margarita; Valêncio, Carlos Roberto y Tenório, William (2021). Corpus-based methodology for an Online Multilingual Collocations Dictionary: First Steps. En I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek y C. Tiberius (Eds.), *Proceedings of eLex 2021* (pp. 1-28).

Padró, Luís y Stanilovsky, Evgeny (2012). FreeLing 3.0: Towards wider multilinguality. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2473-2479. Disponible en: http://hdl.handle.net/2117/15986

Pamies, Antonio (2019). La fraseología a través de su terminología. En J. J. Martín Ríos (Ed.), *Estudios lingüísticos y culturales sobre China* (pp. 105-134). Comares.

Penadés Martínez, Inmaculada (2017). Arbitrariedad y motivación en las colocaciones. *RLA*, *55* (2), 121-142.

Pérez Serrano, Mercedes (2014). ¿Son indispensables los diccionarios combinatorios? *Revista de Lexicografía, 20*, 121–145.

Qian, Xiaofei (2012). Automatic Extraction of Chinese V-N Collocations. En D. Ji y G. Xiao (Eds.), *Chinese Lexical Semantics* (pp. 230-241). Springer. https://doi.org/10.1007/978-3-642-36337-5_24

Qiao, Yun (2017). *Evolución y estructura del léxico chino: Un enfoque cognitivo*. Universidad de Granada.

Santos, Diana y Rocha, Paulo (2005). The Key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. En C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, y B. Magnini (Eds.), *Multilingual Information Access for Text, Speech and Images* (pp. 821-832). Springer. https://doi.org/10.1007/11519645_80

Sinclair, John McHardy (1991). *Corpus, concordance, collocation*. Oxford University Press.

Sinclair, John McHardy (1966). Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday y R. H. Robins (Eds.). *In Memory of J.R. Firth*. Longman.

Straka, Milan y Straková, Jana (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. En J. Hajič y D. Zeman (Eds.), *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 88-99). Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-3009

Sun, Maosong; Huang, Changning y Fang, Jie (1997). Han yu da pei ding liang fen xi chu tan 汉语搭配定量分析初探 ['Un estudio preliminar sobre el análisis cuantitativo de la colocación china']. *Zhong guo yu wen* 中国语文, *1*, 29-38.

Tarp, Sven (2015). La teoría funcional en pocas palabras. *Estudios de Lexicografía*, *4*, 31-42.

Teubert, Wolfgang (2004). Units of meaning, parallel corpora, and their implications for language teaching. En U. Connor y T. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 171–189). Rodopi.

Torner, Sergi y Bernal, Elisenda (Eds.). (2017). *Collocations and Other Lexical Combinations in Spanish*. Routledge. https://doi.org/10.4324/9781315455259

Zeng, Tong (2015). *Ji yu da gui mo yu liao ku de han yu da pei zi dong yan jiu chou qu* 基于大规模语料的汉语搭配自动抽取研究 *['Extracción automática de colocaciones en chino a partir de un corpus a gran escala']*. Nanjing Agricultural University.