Procesamiento del Lenguaje Natural, Revista nº 67, septiembre de 2021, pp. 127-138 recibido 09-05-2021 revisado 07-06-2021 aceptado 09-06-2021

Inducción automática de una taxonomía multilingüe de marcadores discursivos: primeros resultados en castellano, inglés, francés, alemán y catalán

Automatic induction of a multilingual taxonomy of discourse markers: first results in Spanish, English, French, German and Catalan

Rogelio Nazar
Instituto de Literatura y Ciencias del Lenguaje
Pontificia Universidad Católica de Valparaíso, Chile
rogelio.nazar@pucv.cl

Resumen: Este artículo presenta una propuesta metodológica para la inducción automática de una taxonomía multilingüe de marcadores discursivos, que en el caso del castellano corresponden a unidades tales como *sin embargo, por lo tanto, por un lado*, etc. Se propone primeramente un método para separar estas unidades del resto del vocabulario por medio del cálculo de su cantidad de información, seguido de su agrupación en categorías funcionales mediante un corpus paralelo. Finalmente, esta categorización se utiliza como base para la obtención y clasificación de nuevas unidades. Además del método, se describen los primeros resultados, consistentes en una base de datos que actualmente supera ya los 2.600 marcadores.
Palabras clave: inducción de taxonomías, marcadores discursivos, partículas del discurso, lexicografía computacional.

Abstract: This paper presents a methodological proposal por the automatic induction of a multilingual taxonomy of discourse markers which, in the case of English, correspond to units such as *however, therefore, by the way*, etc. First, a method is proposed to separate such units from the rest of the vocabulary using a measure of information, followed by a method to group them using a parallel corpus. Finally, this categorization is used as the basis for the extraction and classification of new units. Apart from the method, the first results are described, which consist of a database that currently surpasses 2600 units.
Keywords: taxonomy induction, discourse markers, discurse particles, computational lexicography.

1 Introducción

Aunque no es un tema nuevo en lingüística, los marcadores del discurso (MD) han estado en el foco de interés de la teoría particularmente en las últimas décadas (Fraser, 1999; Martín Zorraquino y Portolés, 1999; Pons Bordería, 2001, entre otros). Los MD son partículas discursivas que cumplen una amplia variedad de funciones, pero que no forman parte del contenido proposicional de los segmentos a los que afectan. Los ejemplos de estas partículas pueden ser muy diversos, como se explicará más adelante, pero entre los más frecuentes encontramos los conectores aditivos (*además, también*, etc.), los contraargumentativos (*sin embargo, no obstante*, etc.), los causales (*por este motivo, por lo tanto*, etc.) los reformulativos (*es decir, en otras palabras*, etc.), entre un variado número de otras categorías.

La gran mayoría de las investigaciones que se han realizado sobre este tema han sido en el ámbito de la lingüística teórica y con un enfoque cualitativo (cf. Sección 2). Los métodos dominantes hasta ahora han sido la introspección y, en menor medida, el trabajo con corpus. Sin embargo, en este último caso, el corpus es utilizado como herramienta exploratoria, mediante examen visual de líneas de concordancia de uno o algunos MD.

Comparativamente, son pocos los intentos de afrontar este tema con las herramientas del procesamiento del lenguaje natural

ISSN 1135-5948. DOI 10.26342/2021-67-11

(PLN), tanto en castellano como en otras lenguas. La ventaja más evidente del PLN sobre los métodos cualitativos de investigación tradicionales en lingüística en este caso particular es la posibilidad de obtener un inventario masivo de marcadores. Esto es porque, a pesar de corresponder a la categoría de unidades funcionales dentro del vocabulario, no corresponden a una lista cerrada, como la de las preposiciones, y no existe hasta la fecha por tanto un catálogo completo de los MD. Tampoco se ha producido hasta ahora total acuerdo entre los especialistas acerca de cómo se pueden clasificar, ya que los enfoques y teorías son muy diversos y a menudo incompatibles.

El presente artículo pretende hacer un aporte precisamente en la línea del inventariado y la taxonomización de los MD existentes en distintas lenguas. Ofrece una descripción de los resultados preliminares de un proyecto de investigación en curso en el campo de los MD mediante herramientas de PLN. Se trata de una propuesta metodológica para la inducción automática de una taxonomía multilingüe de MD a partir de corpus paralelos, utilizando algoritmos exclusivamente estadísticos. En su estado actual, los resultados del proyecto consisten en una base de datos de 2.636 MD clasificados en 70 categorías funcionales en castellano, inglés, francés, alemán y catalán. Estos datos se encuentran disponibles para descarga desde la web del proyecto[1], y van aumentando en cantidad en la medida en que se continúa con el desarrollo.

La metodología del proyecto incluye una cadena de procesamiento en la que en ningún momento existe intervención humana. Los resultados que se ofrecen, sin embargo, han sido ya revisados por un grupo de lingüistas, hablantes nativos en cada caso, para corregir posibles errores. La tasa de error en los resultados de las diferentes lenguas no superó el 5 % con excepción del alemán, en donde la tasa de error llegó al 16 %.

El método propuesto tampoco utiliza recursos lingüísticos externos tales como vocabularios, diccionarios o etiquetadores morfosintácticos. El único material con el que trabaja es un corpus paralelo de gran tamaño, lo que facilita en gran medida la reproducción de los experimentos en otras lenguas. Como recurso propiamente lingüístico, se utiliza la terminología de Martín Zorraquino y Portolés (1999) para los nombres de las categorías de los MD en castellano, pero esta funciona a modo de metadato externo al propio método y es igual de válida para las distintas lenguas.

Las unidades que el algoritmo inicialmente elige y segmenta como candidatos a MD consisten en palabras o secuencias de palabras consideradas con bajo nivel de información según un cálculo de entropía que se correlaciona con el significado léxico. Según este cálculo, mientras mayor especificidad semántica tiene una palabra, como es el caso de aquellas palabras con una denominación precisa (*aerosol, marxismo, trastorno obsesivo compulsivo*, etc.), mayor es su cantidad de información. Las palabras funcionales o gramemas, tales como las preposiciones, pero también los MD, obtienen según este coeficiente una cantidad de información más baja.

Una vez obtenidos los listados de candidatos a MD, estos son organizados en categorías funcionales a partir del corpus paralelo, explotando su similitud en cuanto a equivalentes en la otra lengua. Esta organización en grupos funcionales se convierte en una clasificación que se realiza, en un primer nivel, con la ayuda de la taxonomía ofrecida por Martín Zorraquino y Portolés (1999), de la que se obtienen los nombres para etiquetar los grupos gracias a los ejemplos que se incluyen. Esta taxonomía, sin embargo, es a su vez subdividida y enriquecida con subcategorías que resultan emergentes del corpus, y que no pueden ser etiquetadas porque exceden el nivel de granularidad de dicho recurso.

Además del interés que puede tener la propuesta en tanto metodología, existe también el que ofrece el resultado mismo. Esto es porque en la bibliografía sobre el tema es frecuente encontrar diferentes taxonomías y listados de ejemplos, pero en la mayor parte de los casos estos alcanzan unos pocos centenares, cuando las unidades utilizadas realmente como MD en la lengua se cuentan por miles. La base de datos que resulta puede tener diversas aplicaciones. Por un lado, puede informar los métodos y las conclusiones de estudios en lingüística teórica sobre el tema. Por otro lado, puede ser utilizado también como herramienta en el PLN para el *parsing* discursivo en tareas de extracción de información. Por último, en su estado actual puede ser también de interés para usuarios finales, ya sea traductores o quienes necesiten redac-

[1] http://www.tecling.com/dismark

tar en su propia lengua o en una L2, y busquen equivalentes o deseen cuidar la riqueza de vocabulario de sus textos.

2 Trabajo relacionado

2.1 Antecedentes teóricos

Entre los pioneros del estudio de los MD se encuentran en particular muchos gramáticos de la lengua castellana, tales como Antonio de Nebrija, Gregorio Garcés, Andrés Bello y más recientemente Gili Gaya (1943), pero la verdadera profusión de investigaciones en el tema es posterior. Comenzó con el trabajo de van Dijk (1973), quien describió las relaciones lógicas que se producen entre proposiciones a través del uso de distintos conectores, tales como los de disyunción, conjunción, causalidad, condición, contraste, etc. Algunos años más tarde, esta línea de investigación se vio extendida por el trabajo de Halliday y Hasan (1976), que presentaron ya una taxonomía más completa para el caso del inglés, incluyendo otras categorías además de las mencionadas por van Dijk. En paralelo, en el área de los estudios de la argumentación en francés, Anscombre y Ducrot (1976) profundizaron en las funciones de partículas y conectores que hoy englobaríamos en la categoría de MD.

Tal como señala Stubbs (1983), el análisis de este tipo de unidades evidenció las limitaciones de lo que hasta los años setenta había sido una gramática oracional y justificó en buena medida el lanzamiento de una gramática del texto, precedente de lo que luego sería el análisis del discurso. A partir de los años ochenta se multiplicaría la cantidad de investigaciones en esta subdisciplina y, particularmente, en el campo de los MD. Los sucesivos trabajos de investigación intentaron delinear las propiedades definitorias de estas unidades, es decir, aquellas que los definen como subconjunto del vocabulario, y también aquellas propiedades que permiten organizarlas en categorías.

Parece existir consenso en que los MD representan un fenómeno común a todas las lenguas, pero no son fácilmente definibles como conjunto de unidades. A menudo son definidos como partículas discursivas que sirven para facilitar las relaciones de coherencia en los textos (Fraser, 1999; Pons Bordería, 2001), en el sentido de que ofrecen instrucciones para la interpretación y van organizando la argumentación. Su aparición, sin embargo, no es estrictamente necesaria ya que igualmente en su ausencia es posible inferir relaciones lógicas entre proposiciones como, por ejemplo, la causalidad. A pesar de que a veces no hacen falta, son sin embargo un elemento clave para facilitar el trabajo interpretativo del lector y reducen el riesgo de error o de ambigüedad.

El rol de los MD también consiste en regular la interacción entre participantes. Esto sucede con mayor frecuencia en la comunicación oral, aunque no exclusivamente. En este sentido, se puede decir que tienen también una función interpersonal además de la textual, o exofórica en lugar de solo endofórica. Mosegaard Hansen (1998), por ejemplo, menciona los indicadores de cambio de tema o de cambio de turno de los participantes en la interacción. Esto hace que consideremos en la categoría de MD a todas aquellas partículas pragmáticas que tienen una función interpersonal, tales como partículas modales e interjecciones, lo que dificulta el establecimiento de un límite preciso.

Desde un punto de vista morfológico, los MD pueden tener diversas categorías gramaticales: conjunciones, adverbios o preposiciones, casi siempre como expresiones pluriverbales. Es posible decir que se caracterizan por ser (relativamente) invariables, ya que no presentan la flexión típica de otros tipos de unidades léxicas. Como explican Martín Zorraquino y Portolés (1999), los MD no presentan flexión de género (**por cierta*) ni de número (**sin embargos*); casi nunca admiten modificadores (**muy sin embargo*, pero sí *muy por el contrario*); no pueden ser negados (**no a saber*) ni coordinados (**a saber y sin embargo*).

Desde un punto de vista sintáctico, Schiffrin (2001) ha señalado que ocupan frecuentemente una posición inicial en la oración, pero también pueden ocupar otras posiciones. Suelen ser también parentéticos, es decir que suelen aparecer entre pausas o, en el caso de la lengua escrita, signos de puntuación, como comas o puntos. Esto parece indicar que no forman parte de la estructura sintáctica de la oración. Sin embargo, nuevamente esta tampoco parece una regla firme, ya que también es posible encontrarlos en una posición no parentética. En cualquier caso, no están confinados a la oración, y tienen la capacidad de afectar alternativamente a distintos niveles, ya sea al intraoracional o bien al extraoracional o discursivo (Pons Bordería, 2001;

Brinton, 2010).

Posiblemente sea el punto de vista semántico el único que permita una distinción más clara del conjunto, ya que se caracterizan por una falta de contenido referencial o proposicional. Aquí también es preciso hacer la salvedad, sin embargo, ya que es posible que algunos conserven parte del significado léxico que alguna vez tuvieron y que perdieron durante la evolución histórica de la lengua a través de un proceso de gramaticalización (Traugott y Dasher, 2002; Wichmann y Chanet, 2009).

Además de los intentos por definir a los MD como conjunto, otro aspecto que ha preocupado a los teóricos es el de diferenciar las distintas clases que existen. En este aspecto, sin duda el trabajo de Halliday y Hasan (1976) es pionero en el esfuerzo de establecer categorías. Sin embargo, nuevamente destaca la tradición española como la que más se ha centrado en la categorización, como se puede apreciar en los trabajos de Casado Velarde (1993), Montolío (2001), Calsamiglia y Tusón (1999) y, en particular, Martín Zorraquino y Portolés (1999).

El último trabajo es el que ha ofrecido la taxonomía más exitosa y que ha influido incluso en la clasificación de MD en otras lenguas, como por ejemplo el alemán (Blühdorn, Foolen, y Loureda, 2017). Consiste en una clasificación en dos niveles: primero ofrece una serie de categorías más generales que luego se subdividen en categorías más específicas. Las categorías más generales coinciden con las que ya han sido señaladas por otros autores, tales como los estructuradores de la información, los conectores, reformuladores, operadores argumentativos y marcadores conversacionales. Pero luego cada una de estas grandes categorías se subdivide y así tenemos entonces, por ejemplo en el caso de los conectores, los aditivos (*además, encima, aparte*, etc.); consecutivos (*por tanto, por consiguiente, por ende*, etc.) y contraargumentativos (*sin embargo, no obstante, en cambio*, etc.).

2.2 Antecedentes de análisis de MD con herramientas de PLN

El tema de los MD ha recibido más atención por parte de la lingüística teórica que de la lingüística computacional o del PLN. En particular, es llamativamente poco tratado en la bibliografía sobre análisis computacional del discurso, que es donde sería más natural encontrarlo. En comparación con el enorme volumen de títulos del área, son pocos los trabajos que tratan explícitamente sobre MD, como Stubbs (1996) o Moore y Wiemer-Hastings (2003).

Además, la gran mayoría de las publicaciones del ámbito de la lingüística teórica dedicada al tema de los MD consiste en el análisis cualitativo de uno o unos pocos casos de MD, como por ejemplo el caso de Urgelles-Coll (2010) en inglés o Cardona (2014) en castellano, entre muchos otros. Comparativamente, son pocos los intentos por ofrecer catálogos exhaustivos de los MD que existen en distintas lenguas, que es justamente el área en la que las herramientas de PLN podrían prestar un mejor servicio. Sí existen algunos esfuerzos por recopilar inventarios amplios de MD, como pueden ser el trabajo de Knott (1996) en el caso del inglés, el de Stede (2002) para el caso del alemán, el de Roze, Danlos, y Muller (2012) para el caso del francés, o los de Santos Río (2003) y Briz, Pons, y Portolés (2008) para el caso del castellano, entre otros. Sin embargo, el esfuerzo humano que exige la compilación manual de estos listados implica una gran dificultad para la obtención de listados verdaderamente exhaustivos. Tal como señalan Lopes et al. (2015), las herramientas de PLN son ideales para esta tarea, y esto puede explicar la aparición de una nueva tendencia en lingüística computacional que descubre un renovado interés por la extracción y catalogación de MD. Y un rasgo común que presentan estos estudios más recientes parece ser el análisis de pares de lenguas, frecuentemente mediante corpus paralelos.

En el caso del citado trabajo de Lopes et al. (2015), el par de lenguas viene dado por la aplicación de un sistema de traducción automática. Parten de un listado de MD en inglés generado de manera manual y se limitan a realizar la traducción de este listado a diferentes lenguas.

En un trabajo anterior (Robledo y Nazar, 2018) se propuso un enfoque basado en clustering a partir en corpus paralelo aplicado al caso de los MD en castellano. Aquel método consistió en obtener grupos de MD con equivalencia funcional, la cual viene dada por compartir equivalentes en otra lengua. La limitación de dicho método es que implica la utilización de variados recursos lingüísticos como etiquetadores morfosintáticos, *gazetteer*

y algoritmos de clustering aglomerativo que son computacionalmente costosos debido a su complejidad cuadrática.

Otros autores han optado por el uso de algoritmos de aprendizaje automático, como Sileo et al. (2019), en el que utilizan como material de entrenamiento un grupo de MD en inglés generado de manera manual. Se concentran en la extracción de MD parentéticos de alta frecuencia y en posición inicial de oración, y el insumo que utilizan son las pistas contextuales, entendidas como enegramas de palabras. También en este caso se trata de una metodología de alta complejidad, tanto conceptual como computacional, que necesita de variados recursos externos que dificultan la reproducción de experimentos en otras lenguas.

En relación con estos esfuerzos recientes para el procesamiento de MD dentro de la lingüística computacional, el presente artículo representa una contribución más en la misma dirección, ya que se propone conseguir un listado amplio de MD. De los trabajos mencionados, el que más se le parece es el de Robledo y Nazar (2018), en tanto explota el uso de corpus paralelos para encontrar la equivalencia entre MD de una misma lengua. En contraste con todos los mencionados trabajos, sin embargo, la virtud principal del que se presenta ahora es que se trata de un método mucho más simple, ya que no requiere prácticamente de ningún recurso externo. Esto representa una gran ventaja en dos sentidos: en primer lugar, disminuye el coste computacional, lo cual facilita el procesamiento de grandes volúmenes de datos, y en segundo lugar, posibilita la reproducción de los experimentos en diferentes lenguas. Finalmente, en contraste con los estudios cualitativos, la ventaja de un enfoque como el que se presenta en este artículo es la gran cantidad de datos que genera, ya que se obtienen listados de miles de MD, en contraste con los pocos centenares a los que llegan la mayoría de los enfoques cualitativos e incluso varios de los que proponen métodos automatizados.

3 Metodología

Como ya se mencionó en la introducción, con esta metodología nos proponemos en primer lugar identificar los MD del corpus separándolos del resto de las unidades del vocabulario (Sección 3.1), para luego clasificarlos de manera inductiva en categorías funcionales (Sección 3.2), que son luego etiquetadas de modo también automático (Sección 3.3). Una vez que existe una taxonomía nuclear o básica, comienza el proceso de poblamiento extensivo de esta estructura (Sección 3.4).

3.1 Vaciado de MD a partir del corpus

La primera fase de la metodología consiste en responder a la pregunta de cómo separar las unidades consideradas MD del resto de las palabras del corpus. Para ello, la decisión fue apostar por una característica propia, aunque no exclusiva, de los MD, que es su bajo contenido informativo.

Naturalmente, no se puede decir que los MD no tengan información en el sentido de que no sean portadores de ningún tipo de significado. Como se mencionó en la Sección 2, los MD poseen un significado funcional, ya que son el vehículo de distintas relaciones de sentido. Pero este es un tipo de significado distinto al valor designador o referencial que tienen típicamente las unidades léxicas. En el extremo de las palabras funcionales encontramos las preposiciones, clase cerrada y perfectamente catalogada en las lenguas conocidas, y en el extremo opuesto los términos especializados. Pero entre un extremo y otro de este continuum encontramos una gran diversidad de unidades que no poseen el significado léxico específico de los nombres o, si lo tuvieron alguna vez, lo perdieron en un proceso de gramaticalización en la historia de la lengua (cf. Sección 2.1).

En este caso, definimos cantidad de información en un sentido formal como un valor que indica cuánto ayuda a predecir una variable aleatoria el resultado de otras variables. Claramente, la distribución de palabras en el corpus no es aleatoria ya que, si lo fuera, la aparición de una palabra no podría informarnos acerca de la aparición de otras. Por ejemplo, si en un texto aparece la palabra *caballo*, existe una probabilidad de que también aparezcan otras palabras de su campo semántico, y esta probabilidad se incrementa cuanto más especializada sea esta unidad. De esta forma, si encontramos una unidad como *trastorno obsesivo compulsivo*, existe una alta probabilidad de encontrar otras que tienen relación con este trastorno, términos de la psiquiatría tales como los síntomas asociados o los fármacos que se utilizan para tratarlo.

No todas las unidades del vocabulario po-

seen esta propiedad, es decir, esta misma cantidad de información, ya que encontramos también palabras en este sentido mucho menos informativas: su aparición en el texto no ayuda a predecir la aparición de otras. Este es el caso de los MD, palabras funcionales cuya aparición no tiene relación con el contenido de los textos en los que aparecen.

Es posible apreciar esta diferencia de manera gráfica. En el primer caso, la Figura 1 muestra la distribución de frecuencias de las palabras que aparecen en los contextos de aparición de *democracia*, en un corpus en castellano, excluyendo gramemas (preposiciones y artículos). Como puede apreciarse, el conjunto de las oraciones que contienen esta palabra contienen también un grupo relativamente amplio de otras unidades que aparecen con alta frecuencia, tales como *humanos, respeto, libertad*, etc. Es en este sentido que decimos que la aparición de la palabra *democracia* nos permite predecir la aparición de otras palabras.

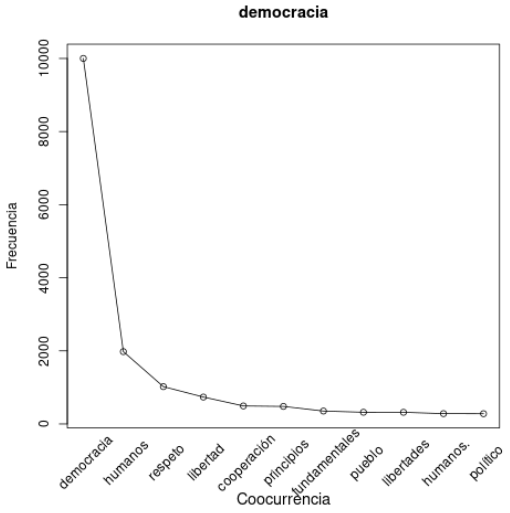

Figura 1: Distribución de frecuencias de las palabras que coocurren con la expresión *democracia*.

Se trata, sin duda, de una propiedad universal del lenguaje, en el sentido de que todas las lenguas ofrecerán un comportamiento similar. No es, sin embargo, el caso de todas las unidades del vocabulario, ya que no será posible predecir qué palabras van a coocurrir con aquellas que tienen un significado funcional en lugar de léxico. En este sentido es que se puede decir que estas palabras tendrán un comportamiento parecido al de una variable aleatoria y, por tanto, su cantidad de información será mucho más baja. Sería el caso de una expresión como *de todas maneras* en el mismo corpus (Figura 2).

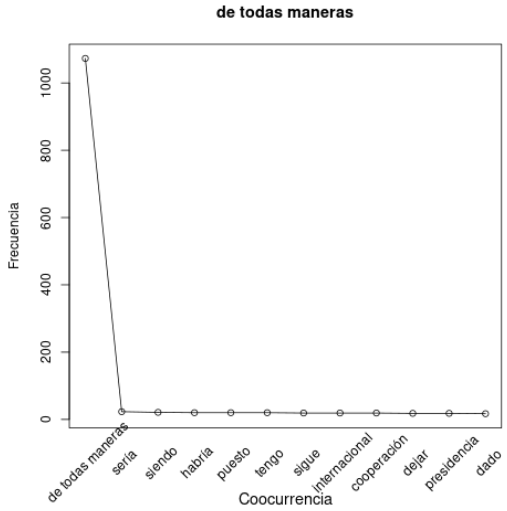

Figura 2: Distribución de frecuencias en el caso de *de todas maneras*.

Comparativamente, las unidades de vocabulario que aparece en las oraciones de unidades funcionales presentan muy baja frecuencia de coocurrencia y son, además, ellas mismas formas poco informativas (*sería, siendo, habría*, etc.). No siempre funcionará esta distinción, ya que hay MD como *por un lado* o *por una parte* que sí permiten la predicción de otras unidades. Pero al menos es posible una primera división del vocabulario en dos clases (palabras informativas vs. palabras no informativas), y los MD genuinos que queden excluidos aquí se podrán recuperar más tarde (apartado 3.4). La división se lleva a cabo utilizando el coeficiente (1), que pone en contraste la suma de las frecuencia de los coocurrentes y la frecuencia de la unidad elegida como diana.

$$I(x) = \frac{\log_2 \sum_{i=1}^{n} R_{x,i}}{\log_2 |m(x)|} \quad (1)$$

Con $m(x)$ nos referimos a los contextos de una unidad x y $R_{x,i}$ es la frecuencia de la unidad i en el *ranking* de los n vocablos más frecuentes en esos contextos (en nuestros experimentos, $n = 20$). Este coeficiente asigna a cada unidad un valor numérico y, por lo tanto, continuo, en lugar de una separación discreta entre dos clases. Ello obliga a elegir un valor de corte arbitrario k para poder establecer la clasificación binaria $C(x)$ (2) entre la categoría léxica (L) y la funcional (F).

$$C(x) = \begin{cases} L & I(x) > k \\ F & \text{otherwise} \end{cases} \quad (2)$$

Para llevar a cabo esta tarea de clasificación, todas las unidades léxicas del corpus deben ser analizadas. Esto requiere la definición de un vocabulario V, en el que $\forall x \in V$, x debe ser una palabra o una secuencia de hasta cuatro palabras. En cuanto al material desde el cual obtener esta información, bastaría con la utilización de un corpus monolingüe lo suficientemente grande como para disponer de unos 5.000 contextos de cada unidad analizada. Sin embargo, como posteriormente vamos a necesitar un corpus paralelo de todos modos, utilizamos para todas las operaciones el mismo corpus, el Opus Corpus ofrecido por Tiedemann (2012).

3.2 Organización en grupos de los MD extraídos

El paso anterior permite obtener, por cada lengua l (en, fr, es, de, ca), un conjunto MD_l de candidatos. El paso siguiente consiste entonces en la agrupación de estas unidades en conjuntos funcionales, para lo cual utilizamos el ya mencionado corpus paralelo.

Es preciso observar aquí algunas de las particularidades del Opus Corpus. Se trata de un conjunto de archivos TMX que se ofrece en pares de lenguas, típicamente en 30 archivos por par, en el que cada uno representa un corpus. Cada corpus reúne material de una determinada área temática o de especialidad, aunque también se encuentra material que corresponde al vocabulario general. Los archivos se encuentran alineados generalmente a nivel de oración. La cantidad total de material disponible varía, por supuesto, según las lenguas elegidas, pero en el caso de las lenguas europeas, cada par está en torno a los 3.500 millones de palabras.

En primer lugar, para poder agrupar los ejemplares de MD obtenidos en el la Sección 3.1, es necesario encontrar los equivalentes de cada uno en otra lengua. Esto es lo que lleva a trabajar por pares de lenguas y, por ende, a la utilización de corpus paralelos. Por una cuestión práctica (la mayor disponibilidad de material) estos pares suelen involucrar al inglés como una de las lenguas, con excepción del catalán, donde tiene más sentido utilizar el par castellano - catalán, que es mayor que el par inglés - catalán. Así, para el caso de un par cualquiera, como por ejemplo castellano - catalán, para la alineación de los conjuntos MD_{es} y MD_{ca} en un listado de equivalentes, utilizamos un coeficiente de asociación basado en un criterio de coocurrencia (3) para encontrar la asociación entre un candidato i en castellano (como, por ejemplo, *en todo caso*) y uno j en catalán (tal como *en tot cas*).

$$A(MD_{es,i}, MD_{ca,j}) = \frac{f(MD_{es,i}, MD_{ca,j})}{\sqrt{f(MD_{es,i})} \cdot \sqrt{f(MD_{ca,j})}} \quad (3)$$

Eventualmente, se podría complementar este coeficiente con otros como el de la similitud ortográfica para el caso de los cognados que son frecuentes en lenguas emparentadas, por ejemplo, nuevamente, el caso del par castellano - catalán. Pero se ha preferido dejar ese recurso de lado para simplificar al máximo el método.

El propósito de alinear los MD en pares de lenguas es únicamente poder agrupar después los MD de una misma lengua en función de los equivalentes que comparten en la otra. De esta manera, se descubrirá la similitud entre dos MD en castellano tales como *en todo caso* y *en cualquier caso* por su mutua relación de equivalencia con un MD en catalán como *en tot cas*. Un aspecto clave de este proceso es que un mismo MD puede ser alineado con distintos equivalentes en otra lengua. Esto sucede con mayor frecuencia en el caso de los MD que en el resto de las unidades léxicas.

Para el descubrimiento de estas relaciones de similitud es preferible evitar el uso de algoritmos de clustering aglomerativo. En lugar de esto, se optó por un método alternativo de mayor simplicidad.

Este nuevo método de clustering está inspirado en las dinámicas sociales que pueden observarse, por ejemplo, en la forma en que se aglutina la gente en las pausas de café de los congresos. Imaginamos un espacio en el que entran personas de a pares, ya que es la situación que tenemos con nuestros MD alineados. El primero puede ser un par cualquiera, como por ejemplo *en todo caso* y *en tot cas*. Si un segundo par que entra no tiene relación con el anterior, entonces permanecen como dos grupos independientes. Sería el caso, por ejemplo, de un par como *en otras palabras* y *en altres paraules*. Ahora bien, si se presenta un tercer par constituido por *en cualquier caso* y *en tot cas*, en ese caso este nuevo par es asimilado el grupo 1, como si *en cualquier caso* fuese presentado a *en todo ca-*

so por *en tot cas*. Esta dinámica continuaría de la misma forma, creando distintos grupos, hasta agotar la cantidad de pares alineados. El proceso resulta económico porque no hay una tabla de distancia en la que se comparen todos los MD entre sí. En cambio, cada par se va comparando con cada uno de los grupos creados hasta el momento. El orden en el que son examinados los pares es aleatorio.

3.3 Etiquetado de los grupos con categorías funcionales

El paso anterior resulta en un número indeterminado de clusters de MD en cada lengua y que se presentan a su vez alineados entre sí. Por ejemplo, el cluster que reúne los conectores contraargumentativos en inglés aparece alineado con el cluster correspondiente en el resto de las lenguas. Estos grupos, sin embargo, no poseen un nombre, tal como suele suceder con el resultado de cualquier proceso de clustering. Esto es, el algoritmo reúne estos conectores por su similitud, pero no los etiqueta con la categoría correspondiente.

Ante este resultado, interesa proporcionar una etiqueta a cada cluster por un criterio lógico de ordenamiento pero también para facilitar el descubrimiento de las relaciones que es posible percibir a simple vista entre algunos clusters. Con este fin, tal como adelantamos ya en la introducción, utilizamos los nombres de categorías aportados por Martín Zorraquino y Portolés (1999). El procedimiento es también aquí bastante simple. Gracias a que estos autores proporcionan varios ejemplos por cada una de estas categorías, es posible encontrar coincidencias (4) entre los miembros de cada una de las categorías en esta taxonomía (MZP) y los miembros de los clusters generados por el algoritmo (CMD).

$$sim(MZP_p, CMD_q) = \frac{|\vec{MZP_p} \cap \vec{CMD_q}|}{|\vec{MZP_p}|} \quad (4)$$

De este modo, para cada cluster se seleccionará la categoría que ofrezca la coincidencia más alta. Naturalmente, como la taxonomía de Martín Zorraquino y Portolés (1999) está en castellano, el cálculo de la intersección solamente puede hacerse con los clusters que están en castellano. Pero esto, por supuesto, no representa un problema debido a que los clusters están alineados interlingüísticamente. De este modo se consigue también el efecto deseado de agrupar clusters que pueden corresponderse a una misma categoría funcional.

3.4 Poblamiento de la taxonomía con nuevos ejemplares

El resultado del paso anterior es una taxonomía multilingüe nuclear o básica, que llamaremos TMD. A partir de este punto, dicha taxonomía puede ser enriquecida mediante la adición de nuevos MD extraídos del corpus. Para cualquier nuevo candidato a MD (c), la existencia de la TMD posibilita decidir si c es efectivamente un MD y, si efectivamente lo es, asignarle una categoría. Para ambas tareas recurrimos nuevamente al corpus paralelo inicial.

Si un candidato c es un MD genuino, entonces su condición será delatada por la presencia de otros MD de la otra lengua en los pares alineados, que ahora es posible descubrir sin dificultad gracias a la taxonomía nuclear. Por ejemplo, si c = *de même façon* $\wedge c \notin TMD$, encontraremos que, en el corpus paralelo francés-inglés, c aparece alineado con elementos tales como *in the same way, likewise, similarly*, etc., elementos que sí aparecen en la TMD. Finalmente, para asignar una categoría a c, operamos de manera similar a 3.3, eligiendo la categoría que ofrece la coincidencia más alta. En el caso del ejemplo, esta correspondería a la de los conectores aditivos.

4 Resultados

En el momento actual, los resultados del proyecto implican la creación de una TMD multilingüe de 2.636 elementos divididos en 70 categorías funcionales. Todavía no ha comenzado el proceso de poblamiento masivo de esta taxonomía, pero sí ha sido posible completar una primera fase de evaluación de la metodología empleada en el proceso. Esta evaluación consiste en medir la capacidad del algoritmo para distinguir entre un MD genuino y una unidad léxica de otra categoría.

La tabla 1 muestra un ejemplo de cluster que corresponde a la categoría de los conectores contraargumentativos según la taxonomía de Martín Zorraquino y Portolés (1999).

Un grupo de lingüistas hablantes nativos de cada una de las lenguas analizadas llevó a cabo una revisión manual de los resultados para evaluar si la selección de marcadores era correcta. Es importante aclarar que lo que se

Inglés	*all the same; although; and yet; but; but still; despite all; despite the fact that; despite these; despite this; even if; even so; even though; however; in spite of all; in spite of the fact; nevertheless; nonetheless; that being said; that said; though; while; yet*
Castellano	*a cambio; ahora bien; al contrario; aparte de eso; a pesar de ello; a pesar de eso; a pesar de esto; a pesar de todo; aun así; aun cuando; aun en; aunque; bien que; con todo; de cualquier forma; de cualquier modo; de todas formas; de todas maneras; de todos modos; dicho esto; en cambio; en lugar de eso; en vez de eso; incluso aunque; no obstante; pero; pero aun así; pese a ello; pese a todo; por el contrario; si bien ; sin embargo; todo lo contrario; y sin embargo*
Francés	*cependant; et pourtant; mais encore; mais toujours; malgré cela; malgré tout; même ainsi; même si ; néanmoins; pourtant; toutefois*
Alemán	*aber immer noch; aber nicht; aber trotzdem; allerdings; auch wenn; auftreten müssen; dachte; dennoch; jedoch; obwohl; selbst wenn; sogar; trotzdem; trotz der tatsache; trotz dieser; trotz dieses*
Catalán	*al contrari; ans al contrari; ben al contrari; de qualsevol manera; de tota manera; de totes formes; de totes maneres; en comptes d'això; en lloc d'això; i no obstant això; malgrat això; no obstant; pel contrari; però tot i així; tanmateix; tot el contrari; tot i així; tot i això*

Tabla 1: Ejemplo de uno de los clusters que corresponde a la categoría de conectores contraargumentativos.

revisó fueron listados de MD fuera de contexto. Esto se debe a que analizar instancias de estas unidades en textos particulares equivaldría a una tarea diferente, ya que una misma unidad puede funcionar como MD en un contexto y en otro no.

La revisión reveló que los datos son de buena calidad, con una pureza en torno el 95 % de media en las distintas lenguas con excepción del alemán, donde la precisión alcanzó el 84 %. Las razones del desempeño inferior en alemán no están del todo claras, pero probablemente puedan estar relacionadas con las características morfológicas de esta lengua. Esto debe continuar estudiándose en trabajo futuro. Otra característica llamativa de los resultados es que en general parece haber una tendencia a tener una cantidad de MD en castellano ligeramente mayor que en las otras, como si esta lengua permitiese mayor diversidad en el uso de estas partículas. Nuevamente, esto debe profundizarse en un estudio contrastivo entre las diferentes lenguas. La presente investigación no ha pretendido, en todo caso, dar respuesta a estos interrogantes sino ofrecer una propuesta metodológica para la obtención de los datos.

En relación con el desempeño general del algoritmo en comparación con otros trabajos mencionados en la Sección 2, es posible afirmar que los resultados obtenidos con el presente método son más numerosos y presentan menor tasa de error. Particularmente en el caso de Robledo y Nazar (2018), que es el más comparable en términos de metodología aunque solo trabajen en castellano, el método presentado aquí es más sensible a los elementos de mediana y baja frecuencia, y la tasa de error a la hora de extraer MD es inferior. Hay que señalar, de cualquier manera, que los objetivos de ambos estudios son distintos. En el caso del estudio anterior se trataba de encontrar categorías de MD. En el presente estudio, en cambio, el foco está puesto en reunir un listado exhaustivo de MD particulares.

Para complementar la evaluación manual general y poner en perspectiva los resultados, invitamos a un grupo de estudiantes avanzados en licenciatura en lingüística a participar de un experimento de evaluación. En total participaron 6 jóvenes, que fueron elegidos entre los que mejores calificaciones obtuvieron en la asignatura de Gramática del Texto, de la Pontificia Universidad Católica de Valparaíso, que trata de manera extensiva el tema de los MD.

Cada estudiante recibió una planilla con 720 unidades en castellano en los cuales se mezclaron MD auténticos con palabras o secuencias de palabras correspondientes a otras diversas categorías. La proporción fue de dos tercios de MD. La instrucción era marcar con un 1 cada unidad que consideraran como MD. No se les permitió consultar diccionarios ni ningún otro recurso y la tarea era individual, sin posibilidad de dialogar con los compañeros. También se les pidió que confiaran en su primera intuición como hablantes, sin dedicar mucho tiempo a cada decisión. La

misma tarea fue realizada por el algoritmo, es decir la de aceptar o rechazar los candidatos del mismo listado. En la Tabla 2 se muestran los resultados de cada uno.

Anotador	**Pre**	**Rec**	**F1**
Algoritmo	97	94	95
Estudiante 1	96	50	65
Estudiante 2	95	60	73
Estudiante 3	95	41	57
Estudiante 4	95	59	72
Estudiante 5	94	65	76
Estudiante 6	92	75	82

Tabla 2: Comparación del desempeño entre algoritmo y humanos en la tarea de separar MD de unidades léxicas (precisión, cobertura y F1).

En general, todos los estudiantes tuvieron un buen desempeño en términos de precisión, en el sentido de que, si seleccionaban una unidad como MD, casi siempre la decisión era correcta. El problema en general es que tuvieron tendencia a ser poco exhaustivos. En comparación con los estudiantes, el algoritmo presentó más o menos la misma tasa de precisión, pero la tasa de cobertura fue mayor.

En una serie de entrevistas realizadas con posterioridad a la entrega del ejercicio, casi todos los estudiantes coincidieron en explicar que adoptaron una actitud conservadora, de modo que ante la duda prefirieron no elegir unidades que, aunque puedan cumplir la función de un MD, no presentan todavía las marcas de los MD prototípicos o que todavía no han finalizado su proceso de gramaticalización. Unidades como *en estas circunstancias* o *en términos más generales*, por ejemplo, fueron rechazadas en la mayoría de los casos a pesar de que en el listado original figuraban como MD auténticos. En otros casos, los estudiantes consultados hicieron referencia a la alta polifuncionalidad (Pons Bordería y Fischer, 2021) de los candidatos inspeccionados, es decir, algunas unidades podrían funcionar como MD solo en algunos casos muy específicos, mientras que en general no tendrían esa función.

Este ejercicio puso de manifiesto el problema de la falta de acuerdo entre los hablantes acerca de lo que es un MD y también la dificultad de tratar con MD fuera de contexto. Más bien, lo propio sería decir que una determinada unidad funciona como MD en un contexto determinado. Esto representa una interesante vía de trabajo futuro pero, nuevamente, trasciende el objetivo de la presente investigación.

5 Conclusiones

Este artículo ha presentado una nueva propuesta metodológica para la extracción automática de una base de datos multilingüe de MD, incluyendo una evaluación de sus primeros resultados. Dicha propuesta es original y, en comparación con trabajos aparecidos recientemente sobre el mismo tema, resulta más simple en términos conceptuales, de dependencia de recursos y en materia de coste computacional. Esto resulta de gran importancia para la reproducción de los experimentos en distintas lenguas.

La base de datos de MD desarrollada hasta el momento se encuentra disponible para su descarga desde la página web del proyecto (cf. nota 1) y, aun tratándose de un trabajo en curso, puede ya servir para múltiples propósitos. Posibles usuarios finales pueden ser traductores o redactores, y posiblemente también docentes de L1 o L2. Los datos pueden ser útiles también para la comunidad del PLN, ya que pueden emplearse para diversidad de tareas vinculadas con el análisis discursivo y la extracción de información.

Muchas tareas han quedado pendientes, como continuar explorando distintas variaciones en la metodología. Esto puede incluir probar con categorías distintas para la clasificación, probar distintos tamaños para la ventana de contexto y hacer un estudio más riguroso del desacuerdo entre anotadores en las distintas lenguas. Otras posibilidades de trabajo futuro serían reproducir experimentos en otras lenguas y, finalmente, una vía que parece atractiva es la de utilizar la taxonomía creada hasta el momento para el descubrimiento de MD polifuncionales.

Agradecimientos

Esta investigación ha sido financiada por el Gobierno de Chile a través del *Proyecto Fondecyt Regular 1191481: Inducción automática de taxonomías de marcadores discursivos a partir de corpus multilingües* (2019-2021). Agradezco a los revisores por sus comentarios y a Irene Renau, por ayudarme a mejorar el artículo en diversos aspectos.

Bibliografía

Anscombre, J.-C. y O. Ducrot. 1976. L'argumentation dans la langue. *Langages*, 42:5–27.

Blühdorn, H., A. Foolen, y Ó. Loureda. 2017. Diskursmarker: Begriffsgeschichte – theorie – beschreibung. ein bibliographischer Überblick. En H. Blühdorn A. Deppermann H. Helmer, y T. Spranz-Fogasy, editores, *Diskursmarker im Deutschen. Reflexionen und Analysen*. Verlag für Gesprächsforschung, Göttingen.

Brinton, L. 2010. Discourse markers. En A. Jucker y I. Taavitsainen, editores, *Historical Pragmatics*. Gruyter Mouton, Berlin.

Briz, A., S. Pons, y J. Portolés. 2008. Diccionario de partículas discursivas del español.

Calsamiglia, H. y A. Tusón. 1999. *Las cosas del decir: manual de análisis del discurso*. Ariel, Madrid.

Cardona, A. L. 2014. *Aproximación funcional a los marcadores discursivos. Análisis y aplicación lexicográfica*. Peter Lang, Frankfurt am Main.

Casado Velarde, M. 1993. *Introducción a la gramática del texto del español*. Arco libros, Madrid.

Fraser, B. 1999. What are discourse markers? *Journal of Pragmatics*, (31):931–952.

Gili Gaya, S. 1943. *Curso superior de sintaxis española*. Minerva, Mexico.

Halliday, M. y R. Hasan. 1976. *Cohesion in English*. Longman, London.

Knott, A. 1996. *A data-driven methodology for motivating a set of coherence relations*. Ph.D. tesis, University of Edinburgh, UK. British Library, EThOS.

Lopes, A., D. M. de Matos, V. Cabarrão, R. Ribeiro, H. Moniz, I. Trancoso, y A. I. Mata. 2015. Towards using machine translation techniques to induce multilingual lexica of discourse markers.

Martín Zorraquino, M. A. y J. Portolés. 1999. Los marcadores del discurso. En *Gramática Descriptiva de la Lengua Española*. Espasa, Madrid, páginas 4051–4214.

Montolío, E. 2001. *Conectores de la lengua escrita. Contraargumentativos, consecutivos, aditivos y organizadores de la información*. Ariel, Barcelona.

Moore, J. D. y P. Wiemer-Hastings. 2003. Discourse in computational linguistics and artificial intelligence. En A. C. Graesser M. A. Gernsbacher, y S. R. Goldman, editores, *Handbook of Discourse Processes*. Routledge.

Mosegaard Hansen, M.-B. 1998. *The Function of Discourse Particles : A study with special reference to spoken standard French*. John Benjamins, Amsterdam/Philadelphia.

Pons Bordería, S. 2001. Connectives/Discourse markers. An Overview. *Quaderns de Filologia. Estudis Literaris*, (6):219–243.

Pons Bordería, S. y K. Fischer. 2021. Using discourse segmentation to account for the polyfunctionality of discourse markers: The case of well. *Journal of Pragmatics*, 173:101–118.

Robledo, H. y R. Nazar. 2018. Clasificación automatizada de marcadores discursivos. *Procesamiento del Lenguaje Natural*, (61):109–116.

Roze, C., L. Danlos, y P. Muller. 2012. Lexconn: a french lexicon of discourse connectives. *Discours - Revue de linguistique, psycholinguistique et informatique*.

Santos Río, L. 2003. *Diccionario de partículas*. Luso-española de ediciones, Salamanca.

Schiffrin, D. 2001. Discourse markers: Language, meaning, and context. En D. Schiffrin D. Tannen, y H. Hamilton, editores, *The Handbook of Discourse Analysis*. Blackwell, Oxford, páginas 54–75.

Sileo, D., T. Van De Cruys, C. Pradel, y P. Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, páginas 3477–3486, Minneapolis, Minnesota, Junio. Association for Computational Linguistics.

Stede, M. 2002. DiMLex: A lexical approach to discourse markers. En A. Lenci y V. D. Tomaso, editores, *Exploring the Lexicon - Theory and Computation*. Edizioni dell'Orso, Alessandria.

Stubbs, M. 1983. *Discourse Analysis. The Sociolinguistic Analysis of Natural Language*. University of Chicago Press, Chicago.

Stubbs, M. 1996. *Text and Corpus Analysis*. Blackwell, Oxford.

Tiedemann, J. 2012. Parallel data, tools and interfaces in OPUS. En *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, páginas 2214–2218, Istanbul, Turkey, Mayo. European Language Resources Association (ELRA).

Traugott, E. y R. Dasher. 2002. *Regularity in semantic change*. Cambridge University Press, New York.

Urgelles-Coll, M. 2010. *The Syntax and Semantics of Discourse Markers*. Continuum, London.

van Dijk, T. 1973. Text Grammar and Text Logic. En *Studies in Text Grammar*. Reidel, Dordrecht, páginas 17–78.

Wichmann, A. y C. Chanet. 2009. Discourse markers: A challenge for linguists and teachers. *Nouveaux cahiers de linguistique française*, 29(4):23–40.