www.revistalogos.cl
Issn Electrónico: 0719-3262

Artículo de Investigación

Detección de marcadores discursivos: el caso de los conectores causal-consecutivos y su polifuncionalidad

Detection of discourse markers: the the case of causal-consecutive connectors and their polyfunctionality

Recibido: 15-01-2022 **Aceptado:** 11-01-2024 **Publicado:** 30-06-2024

Camila Alvarado
Pontificia Universidad Católica de Valparaíso
cam.alvarado.b@gmail.com
0000-0003-2668-1062

Rogelio Nazar
Pontificia Universidad Católica de Valparaíso
rogelio.nazar@pucv.cl
0000-0002-8853-1353

Resumen: Los marcadores discursivos han sido objeto de estudio de numerosas investigaciones en las últimas décadas, y se ha profundizado en su definición y clasificación en el ámbito más general de las partículas, es decir, las partes invariables del discurso. En el presente artículo se describe una propuesta metodológica basada en un enfoque mixto (cuali-cuanti) para la detección automática de marcadores discursivos parentéticos a través de un corpus paralelo inglés-castellano. Proponemos un algoritmo que identifica indicios de una función de marcador discursivo de una unidad y, de ser el caso, su carácter polifuncional. Se presentan resultados de un estudio dirigido a la categoría específica de los conectores causal-consecutivos. El algoritmo en primer lugar determina la eventual condición de marcador discursivo de la unidad, luego señala si pertenece a la categoría específica de conectores causal-consecutivos y, finalmente, señala si hay indicios de polifuncionalidad.

Palabras clave: marcadores discursivos - corpus paralelo - detección automática - conectores causal-consecutivos - polifuncionalidad

Citación: Alvarado, C. & Nazar, R. (2024). Detección de marcadores discursivos: el caso de los conectores causal-consecutivos y su polifuncionalidad. *Logos: Revista de Lingüística, Filosofía y Literatura*, 34(1), 293-308. doi.org/10.15443/RL3412

Abstract: Discourse markers have been the subject of multiple studies in the last decades, and their definition and classification have been deepened in the more general field of particles, that is, the invariable parts of discourse. This article describes a methodological proposal based on a mixed methods approach for the automatic detection of parenthetic discourse markers using an English - Spanish parallel corpus. We propose an algorithm that identifies traces of a discursive marker function of a unit and, if applicable, its polyfunctional character. The results of the study are presented targeting the specific category of consecutive connectors. In the first place, the algorithm defines the eventual discourse marker's condition. Then, it points out whether it belongs to the specific category of consecutive connectors, and, finally, it detects if there are indications of polyfunctionality.

Key Words: discourse markers - parallel corpus - automatic detection - consecutive connectors - polyfunctionality.

1. Introducción

Los marcadores discursivos (MD) han sido objeto de estudio de la gramática/lingüística del texto y los estudios discursivos inicialmente a partir de los años setenta pero de manera más intensa desde los ochenta (De Beaugrande y Dressler, 1997; Martín Zorraquino y Portolés, 1999; Casado Velarde, 2000). De manera progresiva, la gramática fue interesándose por el estudio de los fenómenos extraoracionales, y abordó el texto no como un conjunto de enunciados, sino también un todo coherente y cohesionado, producido en una situación comunicativa definida a partir del cumplimiento del propósito y los objetivos del hablante (De Beaugrande y Dressler, 1997; Cuenca, 2010).

En materia de estructuración de relaciones lógicas dentro del texto, uno de los mecanismos centrales para configurarlas de modo explícito es el uso de MD. El concepto de MD se ha ido transformando a medida que la investigación lingüística lo ha hecho, lo que ha derivado en una variedad de definiciones, propiedades y funciones. De manera generalizada, los MD se definen como unidades lingüísticas que no ejercen funciones sintácticas sino discursivas, ya que tienen la función de guiar las inferencias en la comunicación. Por esto, se reúnen dentro de los enlaces extraoracionales según rasgos

esenciales como la versatilidad en su distribución o la contribución a la coherencia del discurso (Martín Zorraquino y Portolés, 1999).

El conjunto de la bibliografía muestra consenso acerca de una diversidad de propiedades de los MD, tales como su invariabilidad, cristalización o pérdida de capacidad flexiva, pero quedan muchos aspectos por conocer. Entre estos, uno de los más interesantes es cuál puede ser el método para la catalogación exhaustiva y posterior clasificación de los MD existentes en categorías funcionales. Ligado al problema de la clasificación se encuentra, además, el fenómeno de la polifuncionalidad (Hummel, 2012) y, al respecto de esta, cuáles pueden ser los métodos para predecirla o detectarla.

Con miras en estos problemas, hemos planteado una propuesta metodológica para el estudio de los MD en corpus basado en un diseño mixto de investigación (cuali-Cuanti, siguiendo la notación de Creswell, 2009). Nuestra investigación incluye una primera etapa exploratoria de análisis cualitativo de contextos de aparición de marcadores en corpus, y esto trae aparejada la necesidad de acotar el espacio de búsqueda, por lo que en el presente estudio nos centramos en un tipo específico de MD: la categoría de los conectores causal-consecutivos (CCC). El objetivo general es, entonces, detectar automáticamente ejemplos de CCC, señalando además aquellos casos que presenten indicios de polifuncionalidad. Para ello se establecieron además tres objetivos específicos: 1) determinar de manera automática la condición de MD de una expresión en castellano; 2) distinguir entre CCC y otros MD y 3) detectar casos de polifuncionalidad entre los CCC.

El análisis cualitativo que constituye la etapa inicial de este estudio consiste en una exploración manual de casos de conectores en el corpus paralelo para la extracción de una muestra de concordancias de MD en castellano e inglés para obtener una lista inicial de ejemplos de conectores en ambas lenguas. Este listado es utilizado luego en el análisis cuantitativo, donde manejamos por un lado una variable que llamamos índice de variación traductológica (IVT) y mide la estabilidad en la aparición de equivalentes en los segmentos alineados del corpus paralelo y, por otro lado, otras tres variables que serían 1) la condición de MD, 2) condición de CCC y 3) condición de polifuncional.

2. Marco teórico

2.1 Los marcadores discursivos

Tal como ya se anticipó, el estudio de los MD tiene su origen en el cambio de paradigma que se produce en lingüística al pasar de una gramática oracional a una textual. En el ámbito de los estudios del discurso, se entiende a este como práctica social, compleja y heterogénea, que responde a características determinadas a partir de su uso lingüístico contextualizado (Calsamiglia y Tusón, 1999). Bernárdez (1982), después de relevar diversas definiciones del concepto texto, destaca el consenso acerca de que una propiedad central es esta estructuración caracterizada por la coherencia y la cohesión, por un propósito, una situación comunicativa y una finalidad. Se ha destacado, también, la organización de las formas lingüísticas y las relaciones lógicas entre las proposiciones que a su vez se organizan en una macroestructura o esquema global constituido por macroproposiciones, es decir, secuencias lógicas de contenido proposicional organizadas en un nivel de estructuración mayor (Van Dijk, 1978).

Los MD son una pieza fundamental de esta organización macroestructural del discurso. A nivel general, los MD se han definido como unidades funcionales que actúan en un plano extraoracional, es decir, son unidades que ayudan a la conexión de ideas y a la organización discursiva (Martín Zorraquino y Portolés, 1999). En el ámbito de la lexicología, lo anterior corresponde a las dos grandes categorías en que se divide el vocabulario: las unidades léxicas y las funcionales (Escandell, 2007). Las primeras son las que tienen significado léxico y por tanto remiten a conceptos que pueden identificar estados, propiedades, entidades, actividades, etc. Las segundas, en cambio, tienen la función de indicar cómo se combinan los conceptos entre sí, generando conexiones entre proposiciones y guiando las inferencias que surgen durante el proceso comunicativo.

Los MD pertenecen a la categoría de palabras funcionales o gramaticales porque poseen un carácter invariable, funcional y, aunque en algunos se pueda detectar restos de un contenido léxico, de manera general no lo poseen, es decir que han sido vaciadas de contenido por un proceso de gramaticalización (Fuentes-Rodríguez, 2012). Sin embargo, y a pesar de corresponder a esta categoría de palabras, no forman clases cerradas, como sería el caso típico de las preposiciones, compuestas por un número

reducido de elementos que requieren largos procesos históricos para ser modificados. Los MD tampoco pertenecen a la clase abierta, sino que ocupan un espacio intermedio. Aun cuando se proponen como unidades funcionales, los MD son muy numerosos y diversos, son altamente heterogéneos y poseen diversidad de funciones contextuales, interpersonales y textuales (Fischer, 2014). Pueden cambiar con respecto a la posición en que son usados, cambian en la oralidad y la escritura, y están sujetos a infinidad de matices de interpretación según el contexto ya que, en tanto guías de las inferencias del discurso, están intrínsecamente ligados al razonamiento que van haciendo los interlocutores durante el acto de la comunicación.

Los MD se configuran según el discurso en su contexto, es decir, según su organización discursiva. Por lo tanto, son partículas heterogéneas que se regulan en la interacción discurso-contexto. Por ello, cuando un MD es emitido, responde a las relaciones que se generan a partir de los mecanismos que la situación comunicativa permite. En consecuencia, dichos elementos pueden presentar variaciones en su funcionamiento, lo que ha provocado un avance en el campo de las definiciones y clasificaciones de los elementos lingüísticos, tales como los MD, ampliando las limitaciones que podrían presentar estas partículas.

La clasificación de MD también ha sido materia de debate, ya que está naturalmente ligada a su definición, sus funciones y la especificidad de su uso en el discurso. Algunas taxonomías, como la de Martín Zorraquino & Portolés (1999), resaltan las funciones discursivas que desempeñan los MD, ligadas al proceso comunicativo, la organización informativa del discurso, la vinculación semántica y pragmática entre miembros del discurso, la guía de inferencias y las partículas constitutivas de una conversación. Existen otras clasificaciones (Casado Velarde, 1993; Montolío, 2001) que establecen el valor de los significados de los MD a partir de las funciones transoracionales de dichos elementos (por ejemplo, MD con libre desplazamiento en la oración, normalmente entre comas) o que los presentan como mecanismos de cohesión textual entre las oraciones. En cuanto a su estado global, se mantienen las propuestas con respecto a las funciones y a su posible carácter invariable, aunque se reconoce que puede existir multifuncionalidad en dichos elementos.

2.2 Los conectores causal - consecutivos

Contemplada en todas las clasificaciones existentes de MD, la categoría de los conectores ha sido sin embargo definida de maneras distintas según las corrientes teóricas. Como definición general, un conector es un tipo de MD que tiene el propósito de vincular tanto semántica como pragmáticamente miembros del discurso con otros anteriores (Martín Zorraquino y Portolés, 1999). El CCC es un tipo de conector que plantea una relación entre las proposiciones según un orden de causa y consecuencia, donde el miembro anterior refiere a la causa y el miembro discursivo siguiente se presenta como una consecuencia (Martín Zorraquino y Portolés, 1999). Cada uno de estos conectores plantean, por tanto, una relación consecuencial de los elementos del discurso; algunos autores proponen algunos como: *pues, así pues, por tanto, por consiguiente, por ende, de ahí, en consecuencia y de resultas, entonces y así*. Aun cuando todos corresponden a elementos causal-consecutivos, se plantea distintas formas de formular la relación entre las proposiciones discursivas de este carácter. Por un lado, encontramos conectores limitados a mostrar el miembro en el que se encuentran como consecuente del miembro anterior (*pues, así pues*). Por otro lado, encontramos conectores que se utilizan para unir dos proposiciones, es decir, miembro antecedente y consecuente, cuando este es consecuencia de dicho antecedente (*por tanto, por consiguiente, por ende y de ahí*). Finalmente, encontramos también conectores donde el consecuente es un estado de cosas que se producen a partir de otro estado de cosas (Martín Zorraquino y Portolés, 1999).

Siguiendo con la lógica planteada sobre los CCC, estos también han sido definidos como un tipo de MD que tiene la función de generar una relación lógico-semántica a los segmentos textuales, ya sean enunciados o un conjunto de enunciados que residen en relaciones de causa-consecuencia entre la información interconectada (Calsamiglia y Tusón, 1999; Montolío, 2001). Por tanto, desde este conjunto se desprenden los conectores de base causal: conectores causativos y conectores consecutivos. Los primeros responden a la relación de causa entre segmentos textuales y los segundos introducen la consecuencia entre dichos segmentos textuales. Por tanto, este tipo de conector está caracterizado para indicar la conclusión que se deduce de la información entregada por una proposición anterior al conector, explicitando la relación entre los segmentos textuales (Montolío, 2001) y, en consecuencia, se ilustran los rasgos distintivos entre la causalidad y la estructura de consecución para la coherencia del discurso.

2.3 Polifuncionalidad

Los estudios e investigaciones sobre el léxico y el discurso han llevado a variados análisis sobre las funciones y las clasificaciones de las demás clases de palabras, por ejemplo, la polisemia. Dichos estudios han generado que se cuestionen los significados y las clasificaciones sobre las demás clases de palabra, por ejemplo, sobre los MD, ya que las taxonomías anteriores parecen no cubrir todo lo que dichas unidades son capaces de proponer, teniendo en cuenta su capacidad de variar. Los estudios de gramática, semántica, lingüística textual y otras áreas dieron cuenta de que las clases de palabras presentan variaciones y pueden tener distintos significados con el potencial de crear ambigüedad en los contextos y situaciones comunicativas de las que son parte. Esta variación de significado de las clases de palabras, especialmente de las unidades léxicas, podía crear significados distintos pero que aun así conservan algún tipo de relación con el significado inicial. Por lo tanto, si existen dos significados diferentes, pero que se conectan entre sí, existe la polisemia (Escandell, 2007). Esto quiere decir que, a pesar de que las clases de palabras conservan una relación con su función inicial, podrían tener la capacidad de crear una variación o ambigüedad en los significados dependiendo del contexto del discurso.

Los estudios sobre la polisemia han generado el cuestionamiento sobre clases de palabra de corte gramatical como los MD, puesto que las distintas taxonomías no cubren los distintos contextos en que coexisten las unidades funcionales. Este fenómeno ha sido estudiado bajo el nombre de polifuncionalidad (Hummel, 2012; Pardo Llibrer, 2020). Los MD han sido unidades que se han regido por los imperativos del discurso y, por lo tanto, relegados a funciones definidas, excluyéndose de las funciones subjetivas que operan en el discurso y centrándose solo en los procesos de estructuración del discurso (Hummel, 2012). Por tanto, se debe conectar la estructura con las relaciones subjetivas y procedimentales que actúan en los MD, poniendo en cuestionamiento la falta de contenido conceptual con los que estos elementos han sido definidos.

La polifuncionalidad refiere a la multiplicidad de funciones discursivas de un mismo MD (Hummel, 2012; Borreguero y López, 2010) y es por ello similar en esto a la polisemia, ya que se trata de definir interrelaciones semánticas en la polisemia y funcionales en la polifuncionalidad. Por tanto, la clasificación de las funciones de los MD se va a ir construyendo según la situación comunicativa, es decir, las formas en que estos elementos actúen van directamente relacionadas con los propósitos y objetivos del

proceso comunicativo. Por lo mismo, los estudios sobre MD como unidades funcionales estáticas deben expandirse bajo las situaciones de comunicación que se presenten, poniendo el foco en los rasgos que pueden ir desarrollando.

El carácter polifuncional que podrían presentar los MD es, de cierta forma, un proceso de desemantización que permite que un elemento -que puede tener contenido conceptual diluido- pueda desempeñar variadas tareas para articular el discurso (Borreguero y López, 2010). Esto ocurre precisamente por la indeterminación que poseen estos elementos. Por tanto, el carácter polifuncional de los MD va a depender de factores contextuales y, a partir de estos, desempeñar tareas o funciones discursivas que pueden variar, confluir o ser coincidente con más de un significado o valor semántico dentro de una misma situación comunicativa. Además, las clasificaciones de los MD plantean una problemática, ya que no responden al espectro completo de funciones que pueden asumir, puesto que un MD puede llenar diferentes funciones en diferentes contextos (Fischer, 2014). De este modo, las taxonomías que han generado funciones estables relegan a los MD a una sola función predominante pero, a raíz de lo mencionado anteriormente, los MD no pueden ser relegados a una simple función, considerando que existe una amplia gama de funciones que los MD pueden cumplir.

Parece existir consenso en la bibliografía acerca de la inestabilidad de las taxonomías de MD ya que estos presentan una amplia variedad de funciones según el contexto, los interlocutores y las necesidades que debe cubrir según el propósito que debe cumplir. Más que tener asignada una única función, los MD suelen presentar una función dominante y luego una amplia gama de variaciones, en muchos casos sutiles, pero con consecuencias en la interpretación del texto.

3. Metodología

En el presente estudio investigamos una variable de los MD que llamamos índice de variación traductológica (IVT) y que consiste en una medición del grado de estabilidad en los equivalentes que presentan en un corpus paralelo alineado a nivel de oración. Concretamente, queremos determinar si este índice puede servir para predecir las siguientes tres variables: 1) el estatus de MD, 2) el estatus de CCC y 3) la existencia de polifuncionalidad. La investigación se basa, como hemos indicado antes, en un método mixto a partir de una fase cualitativa seguida de una cuantitativa. La primera responde

a una fase exploratoria consistente en el análisis manual de concordancias extraídas del corpus paralelo para la obtención de ejemplos de MC. A partir de este resultado, la segunda fase tiene como propósito medir la asociación entre las variables mencionadas.

Nuestra investigación parte de una hipótesis general y tres subhipótesis. Dado un determinado candidato X_i consistente en una expresión en castellano, tenemos:

- Hipótesis general: La función IVT(X_i) permite determinar si X_i presenta función de MD en el corpus.
- Subhipótesis 1: $X_i \in$ MD si en las oraciones alineadas de X_i predominan MD en inglés.
- Subhipótesis 2: $X_i \in$ CCC si en las oraciones alineadas de X_i predominan CCC en inglés.
- Subhipótesis 3: si $X_i \in$ CCC y además X_i posee un carácter polifuncional, se observará variabilidad en las traducciones acusada por la función IVT(X_i), con presencia de funciones distintas a la predominante.

Como corpus paralelo, en esta investigación se utilizó el corpus castellano-inglés ofrecido en la plataforma *Opus Corpus* (Tiedemann, 2012). Este corpus fue utilizado porque ofrece la opción de seleccionar dos idiomas para generar un corpus paralelo y de extensión variable. La muestra de corpus utilizado fue de un total de dos mil millones de palabras entre castellano e inglés, según el comando Linux *wc,* utilizado para contar las palabras de un corpus. La elección de dicho corpus se justifica por su variedad y cantidad de textos y, por último, mantiene mayor disponibilidad de texto en lengua castellana e inglés. En adelante, designamos a este corpus con la letra *C*.

3.1 Análisis manual de corpus

El análisis cualitativo consiste en el examen manual de corpus paralelo mediante la extracción de concordancias, con el objetivo de obtener una muestra inicial de MD en castellano e inglés que posteriormente se utilizarán en la fase cuantitativa. Para el proceso de extracción de concordancias se realiza una búsqueda de ejemplos de MD en castellano dentro de un corpus paralelo para luego extraer segmentos oracionales alineados en inglés. Dado el extenso tamaño del corpus, para la extracción de concordancias se utilizó un script en el lenguaje Perl (Wall, 1999).

El análisis cualitativo de estas concordancias alineadas revela la función de cada uno de los candidatos según su contexto, y permite generar un listado de equivalentes en inglés. Estos, a su vez, son utilizados en la búsqueda inversa con el objeto de obtener nuevos equivalentes en castellano, lo que permite aumentar progresivamente la muestra inicial de MD en ambas lenguas y clasificarla de acuerdo con funciones. La mayor parte de estos ejemplos corresponde a aquella en la que se enfoca esta investigación, la categoría de CCC, pero también recolectamos instancias de MD en general. El resultado de esta fase es por tanto de dos listados de ejemplos: E_{CCC} corresponde al primero y E_{MD} al segundo, tal que $E_{CCC} \subseteq E_{MD}$. Aquí tenemos, por un lado, en inglés 42 CCC y 175 MD de otras categorías. Por otro lado, en castellano tenemos 55 CCC y 283 MD de otras categorías. En todos los casos estamos hablando de expresiones parentéticas, puesto que a este tipo de MD nos limitamos.

3.2 Detección y clasificación de MD

La segunda fase de análisis consistió en el proceso de detección y clasificación de MD con medios computacionales y cuantitativos. Tal como ya adelantamos, buscamos establecer una relación entre el IVT y las tres variables que se establecen de manera secuencial: la condición de MD, la condición de CCC y la condición de polifuncional.

Dado entonces un determinado candidato X_i, nuestro algoritmo extrae sus contextos de aparición en el corpus paralelo *C* con los segmentos alineados en inglés, un subconjunto de *C* que denominamos C_i. La condición de MD de X_i viene dada por la función $IVT_{MD}(X_i)$ definida en (1):

$$IVT_{MD}(X_i) = \begin{cases} 1 & \frac{|E_{MD} \cap C_i|}{|C_i|+1} > k \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

Se trata del cociente entre la cantidad de veces en que en los segmentos alineados se encontró ya sea un elemento del conjunto E_{MD}, dividido por $|C_i|$ que representa la frecuencia de aparición de X_i en el corpus *C*. La determinación de la condición de MD de X_i se evalúa finalmente mediante una constante *k* que es un parámetro hallado empíricamente.

Si se determina que $X_i \in$ MD, se procede entonces a la siguiente etapa del análisis, que consiste en determinar si $X_i \in$ CCC. Esto se lleva a cabo con la función $IVT_{CCC}(X_i)$ definida en (2):

$$IVT_{CCC}(X_i) = \begin{cases} 1 & \frac{|E_{CCC} \cap C_i|}{|C_i|+1} > q \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

De manera similar al caso anterior, la condición $i \in$ CCC se evalúa por medio de un umbral arbitrario q. En caso de determinarse la pertenencia de X_i al conjunto CCC, se procede a la etapa final del proceso que consiste en la detección de polifuncionalidad. Para esto se propone la función $IVT_{pol}(X_i)$, descrita en (3):

$$IVT_{pol}(X_i) = \begin{cases} 1 & \frac{|E_{CCC} \cap C_i|}{|E_{MD} \cap C_i|} < p \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

En este caso, el coeficiente está basado en la razón entre la cantidad de segmentos obtenidos del corpus (C_i) en los que se encuentran CCC en inglés y la cantidad de segmentos en los que se encuentran MD en general. Como en los casos anteriores, aquí también recurrimos a una constante *p* definida de la misma forma. De esta forma, y ya asumiendo, de nuevo, que $X_i \in$ CCC, el IVT aquí pondera la cantidad de traducciones diferentes a la de CCC que puede tener X_i como la forma de acusar la presencia de polifuncionalidad. En caso de no alcanzarse el umbral, el valor cero significa que no se pronuncia.

4. Resultados

Para evaluar los resultados del método propuesto se trabajó con un listado de prueba consistente en un conjunto de expresiones en castellano que se encuentran frecuentemente entre signos de puntuación, ya sea entre comas o punto y coma. Entre este listado, en adelante llamado *X*, se encuentran 712 unidades de los cuales solo 338 son MD genuinos. Estos últimos corresponden a MD parentéticos como *sin embargo,* mientras que los elementos que no pertenecen a la categoría de MD son elementos como *gracias a Dios*. Cada elemento de este listado fue sometido a examen y, tal como se explica en la sección anterior, se obtuvieron resultados para cada uno en las tres variables contempladas: $IVT_{MD}(X_i)$, $IVT_{CCC}(X_i)$ y $IVT_{pol}(X_i)$. Para la evaluación de los resultados procedemos con las medidas estándar de precisión y cobertura (Manning & Schütze, 1999).

4.1 Evaluación en términos de precisión

Definimos precisión como la proporción de los candidatos correctos entre los que propone el algoritmo, y la estimamos utilizando el listado *X*. A modo ilustrativo, la Tabla 1 muestra un extracto de los resultados.

Tabla 1. Fragmento del listado de resultados

Id	Candidato	Frec.	E_{CCC}	E_{MD}-E_{CCC}	IVT_{MD}	IVT_{CCC}	IVT_{pol}
169	*como resultado de ello*	56	37		1	1	
170	*no sé*	2304	425	299			
171	*de ahora en adelante*	342	40	16	1		
172	*a tal efecto*	1649	237	46	1	1	1
173	*creo que*	36400	5874	2387	1		
174	*a propósito*	1493	228	56			
175	*hasta que*	18212	2185	671			
176	*pobreza*	17037	2995	2398	1		
177	*con mucho*	1914	328	111			
178	*antes bien*	150	16	27	1		
179	*en contraste*	569	102	229	1		
180	*supongo*	1115	227	87			
181	*todavía*	31204	5273	2292	1		
...	...	...	...	...	...	...	...

En esta tabla se puede apreciar, para cada candidato, su frecuencia total en el corpus, seguido de la cantidad de veces en que este candidato muestra como equivalente un elemento CCC; seguido de la cantidad de veces en que fue traducido como un MD de otro tipo; seguido finalmente de las decisiones de cada variante del IVT para determinar el candidato es o no MD; si es o no CCC y, por último, si posee un carácter polifuncional. A un nivel general, el algoritmo analiza de manera correcta a la mayor parte de los candidatos, tal como se observa en la Tabla 2.

Tabla 2. Evaluación de la precisión de los resultados

	Señalados por el algoritmo	**Correctos**	**Precisión**
| X ∩ MD |	361	338	93%
| X ∩ CCC |	106	76	71%
X Polif.	19	9	47%

Lo primero que revelan estos resultados es que es posible distinguir MD con un alto nivel de precisión, pero una vez realizada esta clasificación, resulta más difícil la subclasificación en la categoría de CCC. Por ejemplo, observamos que detecta correctamente la función de MD de una secuencia como *en algunos casos*, pero posteriormente la clasifica de forma incorrecta como CCC, ya que corresponde en realidad a un ordenador. En cuanto a la detección de polifuncionalidad, la tasa de éxito es menor, pero sigue siendo un buen resultado si se tiene en cuenta la amplitud del espacio de búsqueda, ya que señala 19 entre 712 casos y acierta en casi la mitad.

4.2 Evaluación en términos de cobertura

La medida de cobertura viene a indicar, en nuestro caso, cuál es la proporción de MD reales que el algoritmo es capaz de detectar. Para evaluar la cobertura decidimos utilizar un segundo listado conformado únicamente por elementos CCC que se encuentran respaldados por al menos una autoridad en la materia (Casado Velarde, 1993; Martín Zorraquino y Portolés, 1999; Calsamiglia y Tusón, 1999; Montolío, 2001) ya que son mencionados en las clasificaciones propuestas. Esto nos permite una muestra amplia y representativa de la categoría CCC con 55 elementos en total, exhibidos en la Tabla 3. En adelante designamos este listado de referencia como R_{CCC}.

Tabla 3. Muestra de CCC en castellano obtenida de la bibliografía

a consecuencia de ello, a causa, a fin de que, a raíz de ello, así, así pues, así que, como consecuencia, como resultado, como resultado de ello, como resultado de esto, como tal, con el fin de, con el objeto de, con el propósito de, consecuentemente, consiguientemente, dado que, de ahí, de ahí que, de esa forma, de esa manera, de ese modo, de esta forma, de esta manera, de este modo, de manera que, de modo que, de suerte que, de tal modo que, debido a ello, debido a esto, en consecuencia, entonces, para que, por causa, por consiguiente, por dicho motivo, por ello, por ende, por esa razón, por ese motivo, por eso, por estas razones, por este motivo, por esto, por estos motivos, por lo cual, por lo tanto, por tanto, por todas estas razones, porque, pues, y por eso, ya que.

Es importante señalar que el listado R_{CCC} no informa al algoritmo de ninguna manera. Solo se constituye aquí a los efectos de evaluación, que consiste en medir qué proporción de estos CCC nuestro algoritmo es capaz de detectar. Estos resultados se presentan en la Tabla 9.

Tabla 4. Evaluación de la cobertura de los resultados

	Señalados por el algoritmo	**Total**	**Cobertura**
$\lvert R_{CCC} \cap MD \rvert$	51	55	91%
$\lvert R_{CCC} \cap CCC \rvert$	50	55	89%

5. Conclusiones

Esta investigación tuvo como propósito el desarrollo de una metodología para la detección automática de MD, su posterior clasificación en categorías funcionales y finalmente la detección automática de indicios de polifuncionalidad. Como resultado, presentamos una aplicación del método a la categoría específica de CCC. Los resultados muestran que el corpus paralelo puede utilizarse como herramienta para determinar, entre variadas expresiones en castellano, aquellas que cumplen una función de MD. A partir de un análisis y exploración manual inicial del corpus para la recolección de una lista inicial de ejemplos de MD, fue posible utilizar la nueva medida propuesta, IVT, para predecir la condición de MD, de CCC, y la polifuncionalidad de un candidato.

Los resultados obtenidos de este estudio pueden considerarse relevantes, ya que ofrecen una apertura de posibilidades de investigación sobre los contextos de uso que pueden tener las expresiones en castellano, la posibilidad de contrastarlas con su variación traductológica en inglés y, asimismo, realizarlo en distintas lenguas, lo que podría permitir un nuevo contraste entre las expresiones originales y sus alineaciones en cada una de dichas lenguas. Dejamos para trabajo futuro la tarea de refinar el sistema para mejorar las tasas de precisión y cobertura. Además, exploraremos vías para mejorar la eficiencia computacional, tema que no fue estudiado en esta primera aproximación.

Agradecimientos

Esta investigación ha sido posible gracias al financiamiento del Proyecto Fondecyt Regular 1191481: Inducción automática de taxonomías de marcadores discursivos a partir de corpus multilingües (2019-2021).

Referencias bibliográficas

Bernárdez, E. (1982). Introducción a la lingüística. Madrid: Espasa-Calpe S.A.

Borreguero, M., López, A. (2010). *Marcadores del discurso: de la descripción a la definición*. Vol 45. Iberoamericana.

Calsamiglia, H., Tusón, A. (1999). *Las cosas del decir. Manual de análisis del discurso*. Barcelona: Editorial Ariel.

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

Casado, Velarde, M. (2000). Lingüística y gramática del texto: su articulación interdisciplinar. RILCE: Revista de filología hispánica. Vol. 16, pp. 247 - 262.

Creswell, J. (2009). Research design. Qualitative, Quantitative and Mixed Methods Approaches. 3° Edition. SAGE.

Cuenca, J. (2010) Gramática del texto. Madrid: Arco libros.

De Beaugrande, R., Dressler, W. (1997). Introducción a la lingüística del texto. Barcelona: Editorial Ariel S.A

Escandell, M. (2007). Apuntes de semántica léxica. Madrid: UNED.

Fischer, K. (2014). Discourse markers. En Schneider, K., Barron, A. (Ed.) *Pragmatics of discourse.*Vol. 3, pp. 271 - 294. Berlín: De Gruyter.

Fuentes-Rodríguez, C. (2012). Sobre la gramaticalización de los operadores discursivos, como no podía ser de otra manera. LEA. Vol 34. pp. 6 - 32

Hummel, M. (2012). *Polifuncionalidad, polisemia y estrategia retórica. Los signos discursivos con base atributiva entre oralidad y escritura.* Berlín: De Gruyter.

Manning, C., Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

Martín Zorraquino y Portolés, J. (1999). Los marcadores del discurso. En I. Bosque y V. Demonte (dirs.). *Gramática descriptiva de la lengua española* (Vol 3, pp. 4051 - 4214) Madrid: Espasa Calpe.

Montolío, E. (2001). *Conectores de la lengua escrita.* Barcelona: Editorial Planeta S.A

Pardo Llibrer, A. (2020). La polifuncionalidad de los marcadores discursivos en E/LE según unidad y posición. Foro de profesores de E/LE, Volumen(16), 275 – 286. https://doi.org/x10.7203/foroele.0.17090

Tiedemann, J. (2012) Parallel data, tools and interfaces in OPUS. European Language Resources Association (ELRA).

Van Dijk, T. (1978). La ciencia del texto. 3ª edición. Barcelona: Paidos.

Wall, L. (1999). Perl, the first postmodern computer language. https://www.perl.com/pub/1999/03/pm.html/