Propuesta recibida en julio 2023 y aceptada para publicación en noviembre 2023.

Detección de operadores modales: una primera exploración en castellano

Detection of modal operators: first explorations in Spanish

Javier Obreque
Pontificia Universidad Católica de Valparaíso

Rogelio Nazar
Pontificia Universidad Católica de Valparaíso

Resumen

El artículo presenta una propuesta metodológica de carácter mixto — con énfasis en el aspecto cuantitativo — para la detección y registro de operadores modales. Estas unidades pueden definirse como un amplio y heterogéneo conjunto de expresiones que se utilizan en la comunicación escrita y oral para imprimir la visión subjetiva del emisor en su propio enunciado. La presente propuesta se basa en la explotación de un corpus paralelo para aumentar con medios cuantitativos un listado inicial de ejemplos obtenido en una etapa cualitativa. La metodología es simple, efectiva e independiente de lengua, aunque en este primer ensayo nos enfocamos en el castellano.

Palabras clave

modalización, operadores modales, métodos mixtos, corpus paralelo

Abstract

This article presents a mixed methods approach — with emphasis on the quantitative side — for the detection and recording of modal operators. These units are defined as a broad and heterogeneous set of expressions used in written and oral communication to imprint the subjective vision of the writers/speakers in their own utterance. The present proposal is based on the exploitation of a parallel corpus to augment with quantitative means an initial list of examples obtained in a qualitative stage. The methodology is simple, effective, and language independent, although in this first test we focus on the Spanish language.

Keywords

modality, modal operators, mixed methods, parallel corpora

1. Introducción

El estudio de la modalización tiene una larga tradición en las ciencias del lenguaje, que comienza con la lógica de Aristóteles y alcanza su maduración en la filosofía medieval. Los lógicos escolásticos ya distinguían, en una proposición, entre el *dictum* y el *modus* (Ridruejo, 1999). El primer término se refiere al contenido proposicional del enunciado y el segundo especifica un tipo de modalización.

(1) *El gato está sobre la mesa.*

(2) *Parece que el gato está sobre la mesa.*

En el ejemplo (1) se presenta una oración declarativa, descriptiva, sobre un estado de cosas. En (2), en cambio, se expone un enunciado más complejo, en que ese mismo contenido proposicional (*dictum*) aparece subordinado a una marca de modalización (*modus*), y el efecto inmediato es que se comunica el menor grado de certeza del hablante.

El interés de la lingüística por la modalización comenzó a tomar impulso principalmente a partir de la teoría de la enunciación de tradición francesa (Benveniste, 1966, 1974; Kerbrat-Orecchioni, 1987). A partir de estas investigaciones se ha definido la modalización como parte de los rastros de la subjetividad del sujeto productor del discurso y de la situación en la que tiene lugar el acto enunciativo original (Charaudeau, 1994). Se conserva la distinción medieval entre el *dictum*, que presenta hechos o datos con objetividad, y el *modus*, que señala la actitud, posicionamiento, certeza, subjetividad, opiniones, puntos de vista, sentimientos o emociones del emisor, ya sea ante el oyente o ante el contenido de su propio enunciado (Otaola, 1988; Palmer, 2001; Wiebe et al., 2005; Miwa et al., 2012; Taboada, 2016).

En el presente artículo denominamos *operadores modales* (OM) a estas marcas que manifiestan la subjetividad del hablante sobre su enunciado.

DOI: 10.21814/lm.15.2.411

Los OM son un fenómeno universal, ya que al parecer se encuentran en todas las lenguas (Nissim et al., 2013). El problema es que, desde el punto de vista lingüístico, son difíciles de delimitar conceptualmente. No pertenecen a una sola categoría gramatical, ni a una clase cerrada de palabras, ni a unidades puramente funcionales (Blanché, 1966; Lozano et al., 1982; Nuyts, 2005; Müller, 2007; Nuyts, 2016b,a). Se trata, más bien, de una clase abierta que designa un dominio de tipo conceptual y con medios de expresión muy variados, incluyendo elementos poliléxicos (Pottier, 1977; Calsamiglia & Tusón, 1999; Narrog, 2012). Pueden incluirse en este conjunto expresiones tan diversas como *evidentemente, de hecho, debe tener, es posible que, afortunadamente*, etc. Esta heterogeneidad dificulta, por supuesto, la creación y mantenimiento de inventarios exhaustivos (Ruppenhofer & Rehbein, 2012; Marasović et al., 2016; Pyatkin et al., 2021).

El objetivo del presente trabajo es desarrollar un método computacional que permita identificar, extraer y clasificar de manera automática un amplio conjunto de OM de una lengua. Si bien, como decíamos, el inventario completo es muy difícil de obtener, sí tiene sentido al menos intentar el registro de las unidades más frecuentemente utilizadas.

El método que proponemos para ello está basado en el análisis estadístico de un corpus paralelo, lo que implica una mirada interlingüística en que se utiliza como medio de análisis una lengua distinta a la analizada. Otra característica es que, si bien es esencialmente cuantitativo, también puede considerarse de enfoque mixto, ya que incluye una fase cualitativa previa en la que hay un proceso de anotación manual de corpus.

Los resultados preliminares que se muestran en este artículo corresponden a la lengua castellana. Sin embargo, el valor de la investigación está en que, al tratarse de una metodología principalmente cuantitativa, puede ser aplicada a cualquier lengua que disponga de corpus paralelos y de un conjunto inicial de ejemplos de OM, que puede incrementarse gradualmente con la aplicación del mismo método.

El artículo se organiza de la siguiente manera: después de esta introducción, en la sección 2 se presenta un breve marco teórico que explica la definición de los OM como objeto de estudio (2.1), sus categorías principales (2.2), así como la relación entre el registro de OM y la anotación manual de modalizadores (2.3). En la sección 3 se presenta la propuesta metodológica, los resultados en la sección 4 y en la última sección (5) se presentan las conclusiones y el trabajo todavía por realizar. Además, un sitio web[1] acompaña a este artículo, ofreciendo documentación detallada, el código fuente utilizado y los datos de entrada y salida del proceso.

2. Marco teórico

2.1. Operadores modales

Hemos definido los OM como operadores de clase abierta que marcan un tipo de modalización en un enunciado. A continuación, complementamos esta caracterización con una descripción de sus propiedades esenciales.

Una de las propiedades esenciales de los OM es el aumento del poder expresivo y, a la vez, de la complejidad sintáctica y semántica de los enunciados en los que operan (Van Dijk, 1980, 2012; Fuentes Rodríguez, 2003). En la gramática y la lingüística textuales (Bernárdez, 1982; Casado Velarde, 1993; De Beaugrande & Dressler, 1997; Cuenca, 2010), se denomina operador una unidad que actúa de esta manera, es decir, que es exterior a la estructura sintáctica de la cláusula a la que afecta (Fuentes Rodríguez, 2009). Con frecuencia el *modus* es un elemento parentético o, más típicamente, el *dictum* se encuentra en una estructura subordinada.

Algunos autores han identificado a los OM como un tipo de marcador discursivo (Martín-Zorraquino & Portolés, 1999), pero existen algunos argumentos para rechazar esta categorización. Si bien en algunos contextos pueden cumplir ambos roles, un OM en principio no es un marcador discursivo porque su función no es orientar intratextualmente la generación de inferencias que un potencial lector/interlocutor debe hacer durante la comprensión de una unidad comunicativa. La función que cumplen los OM es, en cambio, el resultado de una relación extratextual o exofórica, es decir, la que indica cuál es la posición que toma el sujeto productor del mensaje ante su propio enunciado (Barrenechea, 1979).

En particular, los OM se diferencian de los conectores porque estos, aunque también exceden los límites sintácticos de la oración, actúan como enlace entre dos enunciados diferentes. El OM, en cambio, no conecta sino que, como dijimos, opera en un mismo enunciado (Fuentes Rodríguez, 2009). Nuevamente es necesario aclarar que, en ciertas circunstancias, un OM también puede cumplir funciones de un conector, y viceversa. Este es el caso, por ejemplo, de la expresión *de hecho*, que puede funcionar como un OM, pero también, en algunos contextos, como un co-

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

nector justificativo (Fuentes Rodríguez, 2009) o bien como un operador de refuerzo argumentativo (Martín-Zorraquino & Portolés, 1999). Este sería, sin embargo, solamente un caso de polifuncionalidad (der Auwera & Ammann, 2005).

Los OM han sido objeto de interés de distintas corrientes de la lingüística del texto porque son útiles para la caracterización de los textos. Por ejemplo, algunos investigadores se interesan en medir el tipo de modalización de los textos académicos (Gutiérrez, 2010), describir modalizadores específicos de un determinado género (Sologuren & Venegas, 2022) o describir expresiones modales como marcadores metadiscursivos (Salas, 2015).

2.2. Tipología de operadores modales

Las tres categorías fundamentales de OM proceden de la lógica clásica: epistémica, deóntica y alética. Los OM epistémicos indican el grado de compromiso y conocimiento en los enunciados (Pérez Canales, 2009; González et al., 2016). Tradicionalmente, se han vinculado con los verbos *saber* y *creer* (*pensamos, intuimos, conocemos*), pero también con otras categorías gramaticales, como los adverbios (*aparentemente, supuestamente*, etc.). Los OM deónticos, por su parte, manifiestan la expresión de una obligación (Van Dijk, 1980; Nuyts, 2005) (*se debe, es necesario que*). En cuanto a los OM aléticos, estos expresan necesidad o posibilidad (Van Dijk, 1980) (*es posible que, probablemente*). Esta es una categoría menos utilizada en análisis lingüísticos (Nuyts, 2006), pero está en la base de las dos anteriores (Calsamiglia & Tusón, 1999).

Esta tipología inicial no agota las posibilidades de estudio de las expresiones modales (Narrog, 2012), pero constituye al menos una base útil (Müller, 2007; Portner, 2009). Existen, además, otras categorías que no están del todo asentadas en el análisis del discurso: veredictorias o veredictivas (acerca de la verdad o la mentira de los enunciados), valorativas o axiológicas (evaluación positiva o negativa), volitivas o bulomayeicas (deseo, preferencia o necesidad), de usualidad, de cantidad (Greimas, 1973; Lozano et al., 1982; Otaola, 1988; der Auwera & Plungian, 1998; Calsamiglia & Tusón, 1999; Nuyts, 2006; Cuenca, 2010). A estas pueden sumarse otras incluso menos estables que algunos autores disponen bajo la categoría de indeterminadas (Kalinowski, 1976; Lozano et al., 1982; Portner, 2009).

2.3. Registro de operadores modales: anotación manual y automática

Aunque, como ya se anticipó, la heterogeneidad de los OM hace imposible su consideración como clase cerrada de palabras, hay estudios que se han propuesto identificarlos e inventariarlos (Pyatkin et al., 2021). Existen también recopilaciones enfocadas específicamente en verbos modales (Brandt, 1999; Ruppenhofer & Rehbein, 2012; Marasović et al., 2016, entre otros), es decir, formas verbales que prototípicamente expresan las nociones propias de las modalizaciones del enunciado, como *sé* o *creemos* (modalidad epistémica), *debes* o *haz* (deóntica), etc.

En lengua castellana también se han llevado a cabo esfuerzos destacables por desarrollar inventarios o registros de OM, y tres de ellos corresponden a proyectos lexicográficos de partículas discursivas, como el *Diccionario de partículas* (Santos Río, 2003), el *Diccionario de partículas discursivas del español*[2] y el *Diccionario de conectores y operadores del español* (Fuentes Rodríguez, 2009).

El análisis de los OM también se ha llevado a cabo a través de estudios de corpus (Hendrickx et al., 2012; Nissim et al., 2013; Ghia et al., 2016). En este tipo de estudios la metodología se basa en la anotación manual de corpus, que sirve para incrementar los registros de OM y como recurso para desarrollar propuestas computacionales de detección y anotación de estas partículas (Baker et al., 2010; Rubinstein et al., 2013; Quaresma et al., 2014). Estos registros pueden ser utilizados de forma directa o bien como fuente para crear sistemas de detección basados en reglas (Saurí et al., 2005, 2006; Soni et al., 2014; Lee et al., 2015).

3. Metodología

La presente propuesta metodológica se sustenta en la estrategia de uso de una segunda lengua para el estudio de un aspecto particular de una lengua determinada. Para este primer ensayo, la lengua objetivo es el castellano, y la segunda lengua el inglés, por cuestiones prácticas de disponibilidad de material. En lo esencial, la metodología que describimos en este artículo es la de un proceso de clasificación, en el que a partir de un listado de expresiones (todas las palabras y secuencias de palabras de un corpus), pretendemos clasificarlas en las categorías de OM y no-OM. Según nuestro

[2]Diccionario de Partículas Discursivas del Español, Briz, A. and Pons, S. and Portolés, J. (coords.), `http://www.dpde.es`.

diseño de investigación, se requiere un listado de OM en la segunda lengua, y se requiere aplicar la metodología en las dos direcciones, es decir, invirtiendo la segunda vez el orden de lengua objetivo/medio (castellano/inglés). Además, el proceso debe realizarse de manera independiente por cada categoría de OM. En esta ocasión hemos probado con OM epistémicos, deónticos, aléticos y valorativos. A continuación presentamos los materiales utilizados para esta aplicación metodológica (3.1) y, seguidamente, las fases o etapas del procedimiento (3.2).

3.1. Materiales

El insumo principal de nuestra metodología es un corpus paralelo (CP). Los CP han sido utilizados ya como fuente de información semántica, a modo de espejo para detectar equivalencias (Dyvik, 2004). Esto es, dos elementos en una lengua se pueden considerar similares entre sí cuando el CP revela que tienen los mismos equivalentes en la otra lengua. En el ámbito de los estudios vinculados a expresiones modales ya se han realizado estudios utilizando este recurso, como, por ejemplo, el de Almeida & Carrió Pastor (2015), aunque con un enfoque cualitativo. Desde el enfoque cuantitativo, están relacionados con este estudio los trabajos de Robledo & Nazar (2018) y Nazar (2021), que explotan los CP para extraer marcadores discursivos. La metodología en esos casos es, sin embargo, distinta, ya que se basan en técnicas de *clustering*. La propuesta que exponemos en este artículo no requiere este tipo de algoritmos, que suelen ser computacionalmente costosos. En su lugar, proponemos un método comparativamente más simple, lo cual es preferible según el principio de parsimonia.

El CP utilizado es el Opus Corpus (Tiedemann, 2012), en particular, el subcorpus Scielo de esta colección. Este subconjunto está compuesto por títulos y resúmenes de artículos de investigación de la base de datos del mismo nombre[3]. Esta muestra tiene una extensión de 25.106.776 tokens y fue elegido por corresponder al género argumentativo, terreno fértil para la producción de OM. En el Cuadro 1 se muestra un ejemplo de organización del CP Scielo a partir de la búsqueda del adverbio modal *claramente*.

Como parte de la preparación previa de este material, por cuestión de eficiencia computacional, convertimos el CP del formato TMX original (Figura 1) a un formato TXT en el que en una misma línea se disponen las concordancias alineadas, separadas por un tabulador.

[3] https://scielo.org

3.2. Fases del procedimiento

3.2.1. Creación de un listado inicial de OM en castellano

Se creó un listado inicial de ejemplos de OM del castellano, elaborado inicialmente a partir de los datos de los proyectos Dismark[4] y Text·a·Gram,[5] y aumentado por medio de la anotación manual de un corpus de columnas de opinión, siguiendo los lineamientos de Nissim et al. (2013).

El resultado de esta fase cualitativa inicial fue un listado de 93 OM aléticos (por ejemplo, *podría ser que, con toda probabilidad, es esperable*), 236 epistémicos (*indudablemente, nos consta, claro está*), 142 deónticos (*es fundamental que, necesariamente, urge*) y 188 valorativos (*importantísimo, es favorable, es muy inapropiado*). Nos referiremos a este listado como el conjunto E_m.

3.2.2. Extracción de n-gramas del CP

En cualquiera de las lenguas con las que se trabaje, el material de entrada o input consiste en un listado de vocabulario. En este caso, naturalmente, ese vocabulario procede de una de las lenguas del CP. Se ordenó el vocabulario del CP en listados palabras y secuencias de hasta cinco palabras, definiendo así un conjunto V de n-gramas con $n \leq 5$. V es entonces el conjunto de unidades input ($x \in V$).

Como paso previo del proceso de clasificación aplicamos a V un filtro por medio de un etiquetador morfológico — UDPipe (Straka & Straková, 2017) — para descartar los n-gramas que inician con sustantivo y aquellos que contienen formas verbales con pronombres enclíticos (*promoverse, pautearse, sugerirse*, etc.), ya que son características que no se asocian con los OM.

Un segundo filtro consiste en retener solamente las unidades más frecuentes: las primeras 100.000 en el caso de las monoléxicas, y las primeras 25.000 en el caso de las secuencias de palabras.

3.2.3. Clasificación

Cada elemento x del conjunto V es tomado como input para una función que devuelve un valor binario (*True/False*) para la proposición $x \in OM$. Esta función se basa en la medición de la coocurrencia en el CP entre x y cualquier miembro del ejemplario E_m. Definimos para ello

[4] http://www.tecling.com/dismark
[5] http://www.tecling.com/textagram

	Texto en castellano	**Texto en inglés**
1	Yo percibo eso claramente en el día a día.	I see it clearly in my daily routine.
2	El título lo decía claramente: Arquitectura y negocios.	The title made it clear: Arquitectura y negocios Architecture and Business.
3	Esto se puede observar claramente en la Ilustración 2.	This can clearly be seen in figure 2.

Cuadro 1: Ejemplo de estructura del CP Scielo.

```
    <tu>
      <tuv xml:lang="en"><seg>This is obviously a definition
that incorporates, in this case, sufficient scientific nuance
and, from a lexicographical point of view, minimizes the effect
of circularity by omitting the term dropsy, equivalent to an
accumulation of serous fluid above typical levels.</seg></tuv>
      <tuv xml:lang="es"><seg>Se trata, evidentemente, de una
definición que incorpora, en este caso, matices científicos
suficientes y, desde el punto de vista lexicográfico, minimiza
los efectos de la circularidad al suprimir el término hidropesía
que equivale a una
acumulación de líquido seroso por encima de los niveles
típicos.</seg></tuv>
    </tu>
    <tu>
```

Figura 1: Muestra de segmentos alineados del corpus paralelo en formato TMX.

un conjunto $R(x)$ como el subconjunto de concordancias de la expresión ingresada como input (x) en el CP (1). A partir de la intersección de $R(x)$ con E_m en el CP (2), obtenemos una función $mop(x)$ (3) que mide la coaparición del input x con algún elemento del conjunto E_m en el CP. La decisión se toma por medio de un umbral arbitrario k (4).

$$R(x) = x \cap CP \tag{1}$$

$$int(x) = |R(x) \cap E_m| \tag{2}$$

$$mop(x) = \frac{int(x)}{|R(x)|} \tag{3}$$

$$\forall x \in V, (x \in OM) = \begin{cases} \text{True} & if\ mop(x) > k \\ \text{False} & otherwise \end{cases} \tag{4}$$

En el Cuadro 2 ejemplificamos el proceso en el caso del castellano como lengua objetivo (y, por tanto, con un E_m en inglés) y con x = *claramente*. En la fila 1, la concordancia del CP en inglés no contiene ningún elemento del conjunto E_m (e.g., *predominantly* $\notin E_m$). En la fila 2, en cambio, sí se presenta un caso de intersección de $R(x)$ con elementos del conjunto E_m (*clearly* $\in E_m$). De este modo, el número de veces en que *claramente* aparece en paralelo con algún OM del ejemplario en inglés (E_m) se divide por la cantidad total de ocasiones en que *claramente* aparece en el CP ($R(x)$). Mientras más alta sea esta proporción, mayor la probabilidad de que $x \in OM$.

3.2.4. Repetición del proceso en orden inverso

Si la lengua objetivo es castellano, en el resultado de la primera ejecución es un listado de OM en inglés como, por ejemplo en el caso de epistémicos, $E_m = \{$*certainly, I believe, undoubtedly, ...*$\}$, el paso siguiente consiste en repetir el mismo proceso a la inversa, es decir, utilizar este resultado intermedio en inglés para obtener un conjunto de OM de vuelta al castellano.

4. Resultados

A continuación se describen los resultados de la aplicación de la propuesta metodológica para las categorías de OM aléticos, epistémicos, deónticos y valorativos. Primero presentamos resultados de la aplicación desde un ejemplario del castellano, es decir que los resultados intermedios están en inglés. Posteriormente, presentamos resultados de la segunda aplicación, en el que se utiliza el ejemplario en inglés ahora para obtener el resultado final en castellano. Aunque la investigación considera n-gramas con $n \leq 5$, por limitaciones de espacio solo presentamos algunos ejemplos de tablas de resultados con bigramas y trigramas. También resumimos en gráficas la evaluación del desempeño del algoritmo de detección de OM del castellano en los 5 n-gramas que re-

	CP_o **(castellano)**	CP_m **(inglés)**
1	La producción obtenida se reveló **claramente** internacional, con apenas un trabajo producido en Brasil.	The obtained production was **predominantly** international, with only one study produced in Brazil.
2	Definir **claramente** la cuestión a plantearse.	Define **clearly** the question to be formulated.

Cuadro 2: Ejemplos de no coincidencia (fila 1) y de coincidencia (fila 2) entre OM de ambas lenguas.

N°	**x**	**R(x)**	$int(x)$	$mop(x)$
1	probably related	37	34	89
2	possibly due	67	60	88
3	possibly because	44	39	86
4	possible that	293	251	85
5	probably due	174	149	85
6	generally associated	27	24	85
7	likely that	109	93	84
8	probably because	75	64	84
9	will probably	33	28	82
10	are probably	60	49	80
11	probable that	52	42	79
12	and probably	49	38	76
13	may mean	29	23	76
14	usually occurs	31	24	75
15	and possibly	58	44	74
16	commonly associated	34	26	74
17	this can	384	276	71
18	probably the	73	53	71
19	usually present	27	20	71
20	was probably	46	33	70

Cuadro 3: Muestra de las más altas puntuaciones obtenidas de bigramas en función de la búsqueda de OM aléticos en inglés

N°	**x**	**R(x)**	$int(x)$	$mop(x)$
1	must be the	39	38	95
2	is necessary that	95	91	94
3	are not necessarily	38	37	94
4	necessary that the	42	40	93
5	*the nurse must	26	25	92
6	should be carefully	35	33	91
7	care should be	49	45	90
8	is not necessarily	43	40	90
9	*the physician should	30	28	90
10	must be able	29	27	90
11	necessary to consider	76	69	89
12	there must be	66	60	89
13	professionals should be	46	42	89
14	must be based	46	42	89
15	necessary to know	38	35	89
16	*patients must be	36	33	89
17	*should be informed	28	26	89
18	should be avoided	96	86	88
19	must consider that	26	24	88
20	we must be	25	23	88

Cuadro 4: Muestra de las más altas puntuaciones obtenidas de trigramas en función de la búsqueda de OM deónticos en inglés

presentan nuestros resultados finales. El resto de los datos están en la ya mencionada web del proyecto.

4.1. Resultados intermedios en inglés

A continuación presentamos una muestra con los resultados de los 20 puntajes $mop(x)$ más altos de la aplicación del procedimiento a un listado de bigramas y trigramas en inglés. Los bigramas corresponden, en este caso, a OM aléticos (Cuadro 3) y los trigramas a los OM deónticos (Cuadro 4). Se marcan con asterisco los casos de elementos que no corresponden a un OM. El número total de unidades en inglés es de 512 en el caso de los aléticos, 338 en el caso de los epistémicos, 469 en el caso de los deónticos y 684 en el caso de los valorativos.

4.2. Resultados finales en castellano

La aplicación del procedimiento con V ahora compuesto por listados de n-gramas en castellano con $n \leq 5$ resultó en un total de 1.084 casos de OM (Cuadro 5). Como en el caso anterior, presentamos una muestra con los resultados finales de los 20 puntajes más altos de la aplicación del procedimiento a n-gramas en castellano. El Cuadro 6 presenta bigramas con OM aléticos y el Cuadro 7 con epistémicos. Al igual que en los casos anteriores, se marcan con asterisco los n-gramas mal clasificados.

Para mostrar la evaluación del desempeño del algoritmo de detección de OM en estos resultados finales en castellano, desde las Figuras 2 a la 5 presentamos gráficos de líneas con la precisión acumulada en los primeros 100 candidatos a OM aléticos (Figura 2), epistémicos (Figura 3), deónticos (Figura 4) y valorativos (Figura 5), en

Categoría	**Cantidad**
Aléticos	69
Epistémicos	160
Deónticos	653
Valorativos	202
Total	**1084**

Cuadro 5: Distribución de OM detectados en castellano por categoría

N°	**x**	**R(x)**	$int(x)$	$mop(x)$
1	probablemente debido	51	49	94
2	posiblemente debido	30	29	93
3	probablemente porque	26	25	92
4	pueda causar	9	9	90
5	pudiendo causar	16	15	88
6	nunca será	8	8	88
7	debido posiblemente	7	7	87
8	podría explicarse	28	25	86
9	posible explicación	56	48	84
10	*variar desde	12	11	84
11	podría atribuirse	12	11	84
12	pudiendo llevar	23	20	83
13	*sólo podrá	11	10	83
14	posiblemente porque	16	14	82
15	*explicarse porque	10	9	81
16	* ocurrir después	10	9	81
17	*causar cambios	10	9	81
18	*ocurrir durante	29	24	80
19	pudiendo incluso	9	8	80
20	pudiendo resultar	13	11	78

Cuadro 6: Muestra de las más altas puntuaciones obtenidas de bigramas en función de la búsqueda de OM aléticos en castellano

los cinco tipos de n-gramas ($1 \leq n \leq 5$). El patrón que se observa es que muchos OM se detectan correctamente al inicio y luego la precisión comienza gradualmente a decaer.

4.3. Métricas de evaluación

La precisión de estos resultados finales en castellano fue evaluada cualitativamente por dos anotadores. Para medir el grado de acuerdo se evaluó una muestra aleatoria de 100 casos por cada una de las categorías, constituidas por los distintos tipos de n-gramas en partes iguales. Esta medición arrojó un acuerdo total del 94 % y un índice Kappa de Cohen de 0.89, que puede considerarse alto. La evaluación de la precisión de los resultados fue medida a partir de muestras de 100 casos por cada categoría de OM estudiada, 400 en total. Se seleccionaron muestras de

N°	**x**	**R(x)**	$int(x)$	$mop(x)$
1	aparentemente sanos	18	18	94
2	yo pienso	163	148	90
3	ninguna duda	10	10	90
4	piensa usted	10	10	90
5	claramente definido	9	9	90
6	probablemente porque	26	24	88
7	entonces pienso	8	8	88
8	debido posiblemente	7	7	87
9	lógicamente estabilizado	7	7	87
10	probablemente debido	51	45	86
11	posiblemente debido	30	26	83
12	*usted cree	5	5	83
13	yo creo	200	165	82
14	posiblemente porque	16	14	82
15	queda claro	64	52	80
16	nuestra opinión	53	42	77
17	posible pensar	21	17	77
18	tengo dudas	8	7	77
19	quizá porque	8	7	77
20	resulta claro	8	7	77

Cuadro 7: Muestra de las más altas puntuaciones obtenidas de bigramas en función de la búsqueda de OM epistémicos en castellano

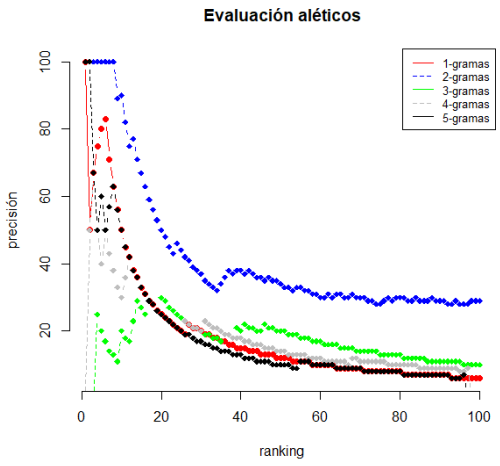

Figura 2: Resultados de la detección de OM aléticos con n-gramas ($1 \leq n \leq 5$)

n-gramas del mismo tamaño a partir de distintas bandas de frecuencia y según la puntuación obtenida: alta, baja y media. Los resultados de este procedimiento se sistematizan en el Cuadro 8. Tal como se puede observar, la mejor precisión del algoritmo se logra en las categorías deóntica (98 %) y epistémica (95 %). En cuanto a los tipos de n-gramas, la mayor precisión del algoritmo se logra con $n = 1$ (93.7 %) y $n = 2$ (96.2 %). El análisis de estos resultados se explicita en la sección 4.4.

n-gramas	aléticos	epistémicos	deónticos	valorativos	T. por n-grama
n = 1	80 %	95 %	100 %	100 %	**93.7 %**
n = 2	95 %	100 %	100 %	90 %	**96.2 %**
n = 3	60 %	95 %	100 %	80 %	**83.7 %**
n = 4	70 %	95 %	100 %	75 %	**85 %**
n = 5	80 %	90 %	90 %	55 %	**78.7 %**
Total	**77 %**	**95 %**	**98 %**	**80 %**	

Cuadro 8: Evaluación de la precisión de detección del algoritmo por categorías de OM y por n-gramas

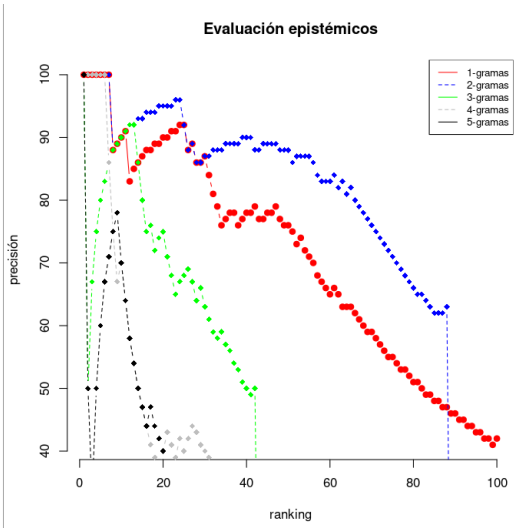

Figura 3: Resultados de la detección de OM epistémicos con n-gramas ($1 \leq n \leq 5$)

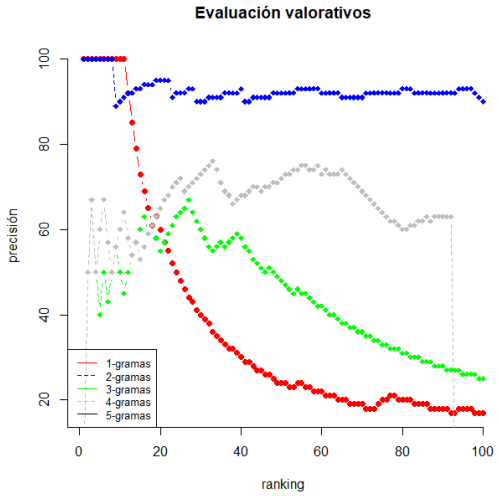

Figura 5: Resultados de la detección de OM valorativos con n-gramas ($1 \leq n \leq 5$)

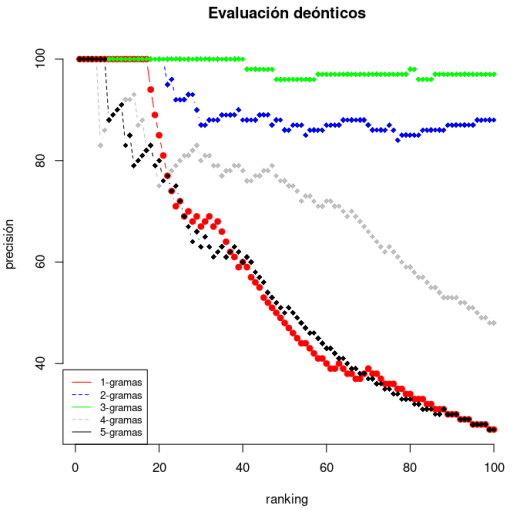

Figura 4: Resultados de la detección de OM deónticos con n-gramas ($1 \leq n \leq 5$)

Como una forma de evaluar la cobertura del método nos inspiramos en el trabajo de Lopes et al. (2015) para desarrollar un *baseline* o método de base. Estos autores se interesan por los marcadores discursivos y aplican un traductor automático para aumentar un listado inicial elaborado manualmente. Para nuestro *baseline*, entonces, imitamos el procedimiento utilizando un traductor automático (DeepL[6]) para convertir el ejemplario inicial en castellano a uno en inglés y utilizar el resultado para traducir de nuevo al castellano. Con este fin, dispusimos los OM del ejemplario inicial en un listado, es decir, uno por línea y seguidos de punto.

El Cuadro 9 muestra los resultados del método propuesto y del *baseline*, y se percibe una diferencia bien marcada. Como se puede observar, en el caso del método propuesto se obtiene mayor diversidad de OM. El traductor automático ofrece sistemáticamente menor variedad. Contrariamente a lo esperado, la intersección entre los resultados obtenidos con nuestro método y los del *baseline* es baja, a tal punto que hace pensar en la posibilidad de combinar ambos métodos en el futuro. En cualquier caso, la intersección con el ejemplario que se muestra en las últimas dos columnas para cada método resume la diferencia en crecimiento con respecto al ejemplario inicial. En todos los casos la comparación favorece a nuestro método por amplio margen, ya que el sistema proporciona más OM y menos coincidencia

[6] https://www.deepl.com/translator

con el ejemplario en comparación con el *baseline*, que con mayor frecuencia acaba reproduciendo los OM del ejemplario inicial.

Otra medida para determinar la cobertura de los resultados obtenidos fue contrastarlos con un listado de 167 OM registrados en el *Diccionario de conectores y operadores del español* (Fuentes Rodríguez, 2009). Esta fuente se eligió porque representa el registro más reciente de estas unidades en castellano. Esta evaluación consideró dos acciones: 1) determinar cuáles de los 167 OM registrados por Fuentes Rodríguez (2009) se encuentran también en el subcorpus Scielo del Opus Corpus y 2) comparar la intersección resultante con los listados obtenidos en esta investigación.

Como resultado de la primera acción se obtuvo que 117 OM registrados en el diccionario de Fuentes Rodríguez (2009) se encuentran también en el corpus Scielo. Como resultado del segundo paso se obtuvo que 15 OM de ese listado fueron detectados por el algoritmo desarrollado en esta aplicación del procedimiento. De acuerdo con esta evaluación, la medida de la cobertura es del 13 %, un número evidentemente bajo. Hay por lo menos dos aspectos para tener en cuenta en la interpretación de este dato.

En primer lugar, los OM registrados por Fuentes Rodríguez (2009) no discriminan entre los que son utilizados en la lengua escrita y en la oral (*a ver, oye, 'eso, eso'*). Aunque algunas de estas unidades pueden encontrarse en un corpus escrito como el de Scielo, su aparición en la escritura no necesariamente está asociada a la expresión de un componente modal de los enunciados, tal como sucedería en el caso de la modalidad oral y, por este motivo, no sería parte de nuestro objetivo detectarlas. En segundo lugar, aunque la intersección de ambos listados es baja, el resultado del registro total de candidatos a OM de las cuatro categorías estudiadas a través de esta propuesta metodológica (1.084 casos) supera ampliamente a los 167 casos registrados por la autora. En esta línea, cabe decir que la medición que hicimos revela que, aun cuando aquí dispongamos de una gran cantidad de OM, debemos concluir que realmente existen muchos más casos por detectar. En este sentido, utilizar CP con otras características, es decir, que incorporen otros géneros discursivos, es una tarea de futuro necesaria para ampliar el registro hasta ahora obtenido.

4.4. Análisis de los resultados

En función de los resultados y su evaluación, relevamos los siguientes aspectos.

En primer lugar, considerando el resumen detallado en el Cuadro 8, se comprueba que la mayor precisión es obtenida en las categorías de OM deónticos y epistémicos. Este resultado podría estar relacionado con que son dos tipos de modalización cuya manifestación lingüística es muy fuerte y, por lo tanto, siempre es muy marcada, incluso dentro de los discursos académicos como los que constituyen el subcorpus Scielo. Este hecho viene a reafirmar la conveniencia de estudiar el fenómeno de la modalización a partir de la división epistémico / deóntica, en tanto categorías modalizadoras fuertes (Müller, 2007; Portner, 2009).

La caída en la precisión de los OM valorativos presentada en el Cuadro 8 desde los n-gramas de 1 y 5 palabras (100 % y 55 %, respectivamente), consideramos que puede estar relacionada con la falta de heterogeneidad semántica, y no solo formal, dentro del ejemplario inicial del proceso (3.2.1). Otra variable de peso podría ser el hecho de tratarse de un CP de lenguaje académico, donde –por estilo y norma– la manifestación de la valoración debe ser restringida. Como resultado preliminar es positivo, pero futuras aplicaciones deberían considerar la posibilidad de nutrir el ejemplario inicial con unidades léxicas provenientes del análisis de sentimientos (Liu, 2010; Zhang et al., 2018, por ejemplo). Además, es probable que a medida que la secuencia de palabras vaya aumentando, su componente modalizador vaya perdiendo fuerza. Estos mismos fenómenos, considerando que no constituyen una marca de manifestación modal fuerte, podrían estar afectando la obtención de una medida de precisión más estable en el caso de los OM aléticos.

Otro aspecto a considerar sería la posibilidad de excluir del análisis las secuencias de más de tres palabras. Esto es porque los resultados muestran que que el *peak* estable de precisión se obtiene en los bigramas (Cuadro 8) y hay motivos para sospechar que el componente modalizador va perdiendo fuerza con el aumento de tamaño de la secuencia de palabras.

Finalmente, a diferencia de los registros existentes, los resultados obtenidos a partir de esta metodología evidencian un aspecto importante sobre la naturaleza de los OM: que no se corresponden ni única ni principalmente con expresiones o construcciones lingüísticas con alto grado de gramaticalización o estabilidad. En esa línea, será necesario un análisis de la estructura formal los OM identificados y de su variación dentro de un mismo paradigma, metodología que excede, por supuesto, los límites de lo planteado en este artículo.

OM	E	M	B	$\|M \cap B\|$	$\|E \cap M\|$	$\|E \cap B\|$
Epistémico	236	160	150	11	23	90
Deóntico	142	653	77	7	13	49
Alético	93	69	71	4	5	51
Valorativo	188	202	152	3	7	112

Cuadro 9: Comparación con un *baseline* (E = ejemplario; M = método propuesto y B = *baseline*)

5. Conclusiones y trabajo futuro

Se ha presentado una propuesta metodológica para detectar OM a partir de un CP con un método principalmente estadístico. En este caso, se han presentado resultados intermedios en inglés que luego fueron utilizados como insumo para la obtención de resultados finales en castellano.

Respecto al desarrollo del algoritmo y los resultados que presentamos, observamos en general un desempeño aceptable, sobre todo en la identificación de OM epistémicos y deónticos.

Una de las limitaciones del estudio es que no podemos ofrecer una comparación con otros métodos más allá del *baseline*, ya que esta es, que sepamos, la primera vez que se propone una tarea de identificación y registro de OM con métodos del procesamiento de lenguaje natural. Es de esperar que futuras propuestas mejoren esta primera aproximación al problema. Otra limitación es que, aunque se ha evaluado la cobertura del método a partir de los OM proporcionados por un diccionario (4.3), sostenemos que esta evaluación pudiera ser parcial u objetable porque no existen, por ahora, registros de OM suficientemente completos en castellano que puedan utilizarse como referencia. En ese sentido, la constitución de un listado de contraste a partir de la anotación manual de un corpus extenso podría ser otra posibilidad que dejamos también para el futuro.

Quedará también para el futuro mejorar el desempeño de este clasificador, en particular cuando se trabaja con n-gramas de $n > 3$. Una posibilidad para ello sería detectar y eliminar palabras en otras lenguas (por ejemplo, *necessariamente*, en portugués) o términos de dominio especializado (por ejemplo, *cianogénicas* o *monoinsaturados*) que a veces se seleccionan por error debido a la naturaleza especializada del subcorpus Scielo. Para ello podría servir un extractor terminológico. También se podría experimentar con CP de otras características, como por ejemplo uno que tenga mayor riqueza expresiva que el discurso académico. Esto, a su vez, redundaría en la obtención de una mayor variedad de OM.

Otros desafíos interesantes para el futuro serán reproducir experimentos con otras lenguas e incluso estudiar los préstamos de modalizadores entre lenguas, fenómeno que ha sido parcialmente explorado por der Auwera & Ammann (2005). También sería interesante estudiar el grado de modulación (alta, media o baja) de un OM, y si existe alguna relación entre este grado y la medida $mop(x)$ de esta propuesta.

Por último, considerando que las categorías analizadas aquí no agotan las posibilidades de estudios de la expresión de la modalización (Narrog, 2012), es parte del trabajo en curso seguir explorando esta metodología con otras categorías (veredictorias, volitivas, de usualidad, entre otras). Ello significará un aporte útil además para evaluar cuáles OM pueden ser polifuncionales entre distintas categorías.

Agradecimientos

El primer autor agradece el apoyo financiero de la Beca de Magíster Nacional/2021 de la Agencia Nacional de Investigación y Desarrollo (ANID) del Gobierno de Chile, que permitió desarrollar esta investigación. Agradecemos también a las revisoras por su trabajo.

Referencias

Almeida, Francisco A. & María. L. Carrió Pastor. 2015. Sobre la categorización de seem en inglés y su traducción en español. análisis de un corpus paralelo. *Revista Signos* 48(88). 154–173. 10.4067/S0718-09342015000200001.

der Auwera, Johan Van & Andreas Ammann. 2005. Modal polyfunctionality and standard average european. En *Modality: Studies in Form and Function*, 247–272. Equinox.

der Auwera, Johan Van & Vladimir A. Plungian. 1998. Modality's semantic map. *Linguistic Typology* 2. 79–124. 10.1515/lity.1998.2.1.79.

Baker, KKathryn, Bloodgood Michael, Bonnie J. Dorr, Nathaniel W. Filardo, Lori Levin &

Christine Piatko. 2010. A modality lexicon and its use in automatic tagging. En *7th International Conference on Language Resources and Evaluation (LREC)*, 1402–1407.

Barrenechea, Ana María. 1979. Operadores pragmáticos de actitud oracional: los adverbios en mente y otros signos. *Estudios lingüísticos y dialectológicos* 39–59.

Benveniste, Émile. 1966. *Problemas de lingüística general I.* Siglo XXI.

Benveniste, Émile. 1974. *Problemas de lingüística general II.* Siglo XXI.

Bernárdez, Enrique. 1982. *Introducción a la lingüística del texto.* Espasa-Calpe.

Blanché, Robert. 1966. *Structures intellectuelles.* Vrin Reprise.

Brandt, Søren. 1999. *Modal verbs in Danish.* C.A. Reitzel.

Calsamiglia, Helena & Amparo Tusón. 1999. *Las cosas del decir: manual de análisis del discurso.* Ariel.

Casado Velarde, Manuel. 1993. *Introducción a la gramática del texto en español.* Arco Libros.

Charaudeau, Patrick. 1994. *Grammaire du sens et de l'expression.* Hachette.

Cuenca, María Josep. 2010. *Gramática del texto.* Arco/Libros.

De Beaugrande, Robert A. & Wolfgang. U. Dressler. 1997. *Introducción a la lingüística del texto.* Ariel.

Dyvik, Helge. 2004. Translations as semantic mirrors: from parallel corpus to wordnet. En *23rdrd International Conference on English Language Research on Computerized Corpora (ICAME)*, 309–326. 10.1163/9789004333710_019.

Fuentes Rodríguez, C. 2009. *Diccionario de conectores y operadores del español.* Arco/Libros.

Fuentes Rodríguez, Catalina. 2003. Operador/conector, un criterio para la sintaxis discursiva. *Rilce* 19(1). 61–85. 10.15581/008.19.26730.

Ghia, Elisa, Lennart Kloppenburg, Malvina Nissim, Paola Pietrandrea & Valerio Cervoni. 2016. A construction-centered approach to the annotation of modality. En *12th ISO Workshop on Interoperable Semantic Annotation*, 67–74.

González, Ramón, Dámaso Izquierdo & Óscar Loureda. 2016. *La evidencialidad en español: teoría y descripción.* Iberoamericana Vervuert. 10.31819/9783954878710.

Greimas, Algirdas J. 1973. Les actants, les acteurs et les figures. En *Sémiotique narrative et textuelle*, 161–176. Larousse.

Gutiérrez, Rosa M. 2010. Especialización del discurso: Una caracterización desde el sistema de la obligación. *Revista de Lingüística Teórica y Aplicada* 48(1). 105–132. 10.4067/S0718-48832010000100006.

Hendrickx, Iris, Amália Mendes & Silvia Mencarelli. 2012. Modality in text: a proposal for corpus annotation. En *8th International Conference on Language Resources and Evaluation (LREC)*, 1805–1812.

Kalinowski, Georges. 1976. Un aperçu élémentaire des modalités déontiques. *Langages* 43. 10–18.

Kerbrat-Orecchioni, Catherine. 1987. *La enunciación: de la subjetividad en el lenguaje.* Edicial.

Lee, Kenton, Yoav Artzi, Yejin Choi & Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1643–1648. 10.18653/v1/D15-1189.

Liu, Bing. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing* 627–666.

Lopes, António, David Martins, Vera Cabarrão, Ricardo Ribeiro, Helena Moniz, Isabel Trancoso & Ana Isabel Mata. 2015. Towards using machine translation techniques to induce multilingual lexica of discourse markers. ArXiv [cs.CL]. 10.48550/arXiv.1503.09144.

Lozano, Jorge, Cristina Peña-Marín & Gonzalo Abril. 1982. *Análisis del discurso: Hacia una semiótica de la interacción textual.* Cátedra.

Marasović, A., M. Zhou, A. Palmer & A. Frank. 2016. Modal sense classification at large: Paraphrase-driven sense projection, semantically enriched classification models and cross-genre evaluations. *Linguistic Issues in Language Technology (LiLT)* 14. 1–58.

Martín-Zorraquino, M. Antónia & José Portolés. 1999. Los marcadores del discurso. En *Gramática descriptiva de la lengua española*, vol. 3, 4051–4213. Espasa. 10.15581/008.16.27325.

Miwa, Makoto, Paul Thompson, John McNaught, Douglas B. Kell & Sophia Ananiadou. 2012. Extracting semantically enriched events from biomedical

literature. *BMC Bioinformatics* 13. 108. 10.1186/1471-2105-13-108.

Müller, Gisela. 2007. Metadiscurso y perspectiva: Funciones metadiscursivas de los modificadores de modalidad introducidos por 'como' en el discurso científico. *Revista Signos* 40(64). 357–387. 10.4067/S0718-09342007000200005.

Narrog, Heiko. 2012. *Modality, subjectivity, and semantic change: A cross-linguistic perspective*. Oxford University Press.

Nazar, Rogelio. 2021. Inducción automática de una taxonomía multilingüe de marcadores discursivos: primeros resultados en castellano, inglés, francés, alemán y catalán. *Procesamiento del Lenguaje Natural* 67. 127–138.

Nissim, Malvina, Paola Pietrandrea, Andrea Sansó & Caterina Mauri. 2013. Cross-linguistic annotation of modality: a data-driven hierarchical model. En *9th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, 7–14.

Nuyts, Jan. 2005. The modal confusion: on terminology and the concepts behind it. En *Modality: Studies in Form and Function*, 5–38. Equinox.

Nuyts, Jan. 2006. Modality: Overview and linguistic issues. En *The Expression of Modality*, 1–26. De Gruyter. 10.1515/9783110197570.1.

Nuyts, Jan. 2016a. Analyses of de modal meanings. En *The Handbook of Modality and Mood*, 31–49. Oxford University Press.

Nuyts, Jan. 2016b. Surveying modality and mood: An introduction. En *The Handbook of Modality and Mood*, 1–8. Oxford University Press.

Otaola, Concepción. 1988. La modalidad (con especial referencia a la lengua española). *Revista de Filología Española* 68(1/2). 97–117. 10.3989/rfe.1988.v68.i1/2.414.

Palmer, Frank R. 2001. *Mood and modality*. Cambridge University Press. 10.1017/CBO9781139167178.

Pérez Canales, José. 2009. *Marcadores de modalidad epistémica: un estudio contrastivo (francés-español)*: Universitat de València. Tesis Doctoral.

Portner, Paul. 2009. *Modality*. Oxford University Press.

Pottier, Bernard. 1977. *Lingüística general: teoría y descripción*. Gredos.

Pyatkin, Valentina, Shoval Sadde, Aynat Rubinstein, Paul Portner & Reut Tsarfaty. 2021. The possible, the plausible, and the desirable: Event-based modality detection for language processing. ArXiv [cs.CL]. 10.48550/arXiv.2106.08037.

Quaresma, Paulo, Amália Mendes, Iris Hendrickx & Teresa Gonçalves. 2014. Automatic tagging of modality: identifying triggers and modal values. *Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation* 95–101.

Ridruejo, Emilio. 1999. Modo y modalidad. El modo en las subordinadas sustantivas. En *Gramática descriptiva de la lengua española*, vol. 2, 3209–3252. Espasa.

Robledo, Hermán & Rogelio Nazar. 2018. Clasificación automatizada de marcadores discursivos. *Procesamiento del Lenguaje Natural* 61. 109–116. 10.26342/2018-61-12.

Rubinstein, Aynat, Hillary Harner, Elizabeth Krawczyk, Daniel Simonson, Graham Katz & Paul Portner. 2013. Toward fine-grained annotation of modality in text. En *IWCS workshop on annotation of modal meanings in natural language*, 38–46.

Ruppenhofer, Josef & Ines Rehbein. 2012. Yes we can!? Annotating English modal verbs. En *8th International Conference on Language Resources and Evaluation (LREC)*, 1538–1545.

Salas, Millaray. 2015. Una propuesta de taxonomía de marcadores metadiscursivos para el discurso académico-científico escrito en español. *Revista Signos* 48(87). 95–120. 10.4067/S0718-09342015000100005.

Santos Río, Luis. 2003. *Diccionario de partículas*. Luso-Española de Ediciones.

Saurí, Roser, Robert Knippen, Marc Verhagen & James Pustejovsky. 2005. Evita: a robust event recognizer for QA systems. En *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 700–707.

Saurí, Roser, Marc Verhagen & James Pustejovsky. 2006. Annotating and recognizing event modality in text. En *19th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, 333–338.

Sologuren, Enrique & René Venegas. 2022. Marcadores epistémicos en el género trabajo final de grado en español: variación disciplinar en la escritura de formación académica. *Literatura y lingüística* 45. 235–258. 10.29344/0717621X.45.2200.

Soni, Sandeep, Tanushree Mitra, Eric Gilbert & Jacob Eisenstein. 2014. Modeling factuality judgments in social media text. En *52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 415–420. 10.3115/v1/P14-2068.

Straka, M. & J. Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. En *CoNLL Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99. 10.18653/v1/K17-3009.

Taboada, Maite. 2016. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics* 2. 325–347. 10.1146/annurev-linguistics-011415-040518.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. En *8th International Conference on Language Resources and Evaluation (LREC)*, 2214–2218.

Van Dijk, Teun A. 1980. *Texto y contexto: Semántica y pragmática del discurso*. Cátedra.

Van Dijk, Teun A. 2012. *Discurso y contexto*. Gedisa.

Wiebe, Janyce, Theresa Wilson & Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation* 39(2). 165–210. 10.1007/s10579-005-7880-9.

Zhang, Lei, Shuai Wang & Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8(e1253). 10.1002/widm.1253.