Marcadores discursivos en árabe y español: un estudio computacional basado en corpus paralelos con anotación pragmática

Ana González-Ledesma y Doaa Saamy

Universidad Autónoma de Madrid y Universidad del Cairo
Laboratorio de Lingüística informática y Dpto. de Filología Española
Carreta de Colmenar Viejo Km. 25 y Faculty of Arts, Main Campus, Cairo University 12613 Giza, Egypt.
ana.ledesma@lllf.maria.com y dsamy@cu.edu.eg

Abstract

El objetivo de este artículo ha sido el de analizar el modo en que se ha llevado a cabo la traducción de los marcadores discursivos del árabe y del español en el corpus paralelo de la ONU desde una perspectiva computacional. La investigación está dividida en tres partes. La primera de ellas está dedicada a la presentación de los recursos. En ella se exponen las características más importantes del corpus de la ONU, por un lado, y por otro, se explica el modelo de anotación pragmática (PRAGMATEXT) utilizado para clasificar los marcadores discursivos. Los fenómenos de naturaleza semántico-pragmática que se explican en el modelo de anotación son: lenguaje emocional, relaciones discursivas, actos de habla, modalización, evidencialidad y deixis. La segunda parte está dedicada a los marcadores discursivos en la parte española del corpus. En ella se explicarán los fenómenos discursivos que se han codificado a través de los marcadores discursivos, así como la frecuencia de aparición de estos últimos. Finalmente, en la tercera parte se explicarán las estrategias computacionales realizadas para la localización automática de marcadores discursivos en el corpus árabe a partir de la información anotada en el corpus español. Por último, se expondrán los resultados de dichas estrategias y se aludirá, en las conclusiones, a algunas de las utilidades de los corpus paralelos anotados con información pragmática.

This article presents an analysis of the translation of Discourse Markers in a parallel Spanish-Arabic corpus from a computational perspective. The research carried out is divided into three main sections. The first section describes the resources used in the study including the main characteristics of the corpus and the pragmatic annotation model (PRAGMATEXT) used in tagging and classifying the discourse markers. The annotation model addresses six semantic-pragmatic phenomena: emotional language, discursive relations, speech acts, modalization, evidentiality and deixis. In the second section, discourse markers in the Spanish corpus are analyzed in detail pointing out the discursive phenomena encoded through the different markers and their frequencies in the corpus. Finally, the third section is dedicated to the technical and computational strategies adopted for the detection of equivalent discourse markers in the aligned Arabic corpus based on the pragmatic information tagged in the Spanish corpus, followed by an evaluation of these strategies. Conclusions regarding the benefits of building resources enriched with pragmatic information are discussed at the end.

Palabras clave: corpus paralelos árabe-español, anotación pragmática, reconocimiento y clasificación semiautomática de marcadores discursivos.

Keywords: parallel Spanish-Arabic corpus, pragmatic annotation, semiautomatic classification of discourse markers.

Tabla de contenidos

0. Introducción

Presentamos en este trabajo un estudio de carácter interdisciplinar donde participan diversas áreas de conocimiento: la Pragmática Lingüística, la Traducción, la Lingüística Contrastiva apoyada en Corpus Paralelos y la Lingüística Computacional.

Es un hecho conocido que la Pragmática, desde sus diferentes marcos teóricos, ha puesto de manifiesto la importancia que los marcadores discursivos tienen a la hora de guiar las inferencias del interlocutor durante el proceso de interpretación de enunciados (Portolés 2004). Poco a poco, el mundo de la traducción está incorporando en su haber el conocimiento generado en esta disciplina, fundamentalmente en lo que se refiere al estudio de la cortesía, los actos de habla, la modalización discursiva y la coherencia y cohesión textual (Hackey 1998). A su vez, la Lingüística Contrastiva ha ampliado en los últimos tiempos sus horizontes, y cada vez son más los estudios comparativos sobre los marcadores del discurso en diversas lenguas. En este sentido, el recurso a los corpus paralelos como material de estudio ha revitalizado con creces esta disciplina (Granger 2003).

Al igual que en otras parcelas, la Lingüística Computacional y la Inteligencia Artificial están trabajando para incluir en sus modelos de lenguaje las relaciones discursivas. En estas parcelas del saber de naturaleza más aplicada, los marcadores discursivos se intentan reconocer y clasificar (de la manera más automatizada posible) con el fin de, bien segmentar el texto automáticamente, bien inducir una estructura retórica. Dentro de este marco, una de las investigaciones que más frutos ha aportado a diferentes campos del mundo computacional, como por ejemplo los resúmenes automáticos, ha sido la aplicación del modelo de relaciones retóricas SDRT propuesto por Mann y Thompson (1988) e implementado computacionalmente en los trabajos de Daniel Marcu (Marcu 2000). En España, los marcadores discursivos también han sido un instrumento a la hora de segmentar textos y de inducir una retórica textual en los corpus (Alonso 2002 y Prada 2003) con fines sobre todo al resumen automático.

No obstante, a pesar de estás iniciativas, consideramos que, fundamentalmente en lo que concierne a la lengua española, y desde una perspectiva computacional, los marcadores discursivos no han sido tratados en toda su complejidad. Por un lado, nos encontramos trabajos procedentes de la Pragmática lingüística que insisten en la polifuncionalidad de estas partículas; estos trabajos se muestran muy reticentes a la hora de presentar una definición definitiva y única del valor semántico-pragmático de un marcador, ya que insisten en que el significado último de los mismos se construye con ayuda del contexto en el momento de la interacción comunicativa. Desde la perspectiva computacional, en cambio, los problemas de ambigüedad, tanto categorial como discursiva, como los criterios

y debates sobre la clasificación semántica de cada marcador no suelen exponerse con detalle en los trabajos consultados.

Por lo que respecta al árabe, el estudio de los marcadores del discurso por medio de corpus ha recibido escasa atención en el ámbito académico europeo, tanto desde una perspectiva teórica como computacional.

Acogiéndonos a estas líneas de investigación en curso, e intentando suplir en parte las carencias señaladas, presentamos aquí un estudio en el que analizamos cómo se han traducido los marcadores discursivos en el corpus paralelo árabe-español de la ONU.

En cuanto a la estructura del discurso, hemos diferenciado los siguientes apartados. En primer lugar presentamos, en dos apartados diferentes, los recursos con los que partimos para desarrollar la investigación, a saber: el corpus paralelo de la ONU y el modelo de anotación pragmática PRAGMATEXT. En el primer apartado se explicarán, brevemente, las características fundamentales de diseño y composición del corpus, así como el tipo de información lingüística que está explicitada. En el segundo apartado, presentamos el marco teórico en el que se fundamenta nuestra clasificación de marcadores discursivos y su formalización a lenguaje XML. Una vez explicados los recursos disponibles, nos adentramos en la parte del procesamiento computacional del corpus paralelo árabe-español de la ONU. Siguiendo con el orden establecido, en el tercer apartado hablaremos de los marcadores discursivos encontrados en el corpus y de su frecuencia de uso. El cuarto apartado está dedicado al procesamiento del corpus del árabe a partir de la información extraída del corpus del español. En él, se expondrán las técnicas computacionales que hemos utilizado para el reconocimiento y etiquetado automático de los marcadores discursivos en el corpus en árabe. A continuación, en el quinto apartado el lector podrá consultar las frecuencias de uso de los marcadores en utilizados en el corpus árabe y a algunos comentarios sobre la evaluación de las estrategias de anotación. Finalmente, para terminar y como es de rigor, se presentarán las conclusiones y el trabajo futuro.

1. El corpus de la ONU: diseño y características

Para este estudio, hemos utilizado un corpus paralelo bilingüe español-árabe, formado por una colección de textos disponibles en Internet procedentes de los documentos de la Organización de las Naciones Unidas.

Términos como "corpus paralelo", "corpus de traducción" y "corpus comparable" pueden resultar ambiguos en algunos casos; por ello, conviene destacar cuál ha sido nuestra concepción de corpus paralelo en el presente trabajo. Desde la perspectiva de la **Lingüística Computacional y el Procesamiento del Lenguaje Natural**, el término "corpus paralelo" sirve para denominar a dos conjuntos de textos en dos lenguas diferentes, los textos de la L2 son traducciones de los textos de la L1. En cambio, el concepto de un "corpus comparable" o un "corpus de traducción" es utilizado para referirse a un conjunto de textos T1 en una lengua A y a un conjunto de textos T2 en una lengua B comparables según el género y la temática. En el presente estudio, empleamos el término de corpus paralelo según la definición de Somers (2001):

> By 'parallel corpus', we mean a text which is available in two (or more) languages: it may be an original text and its translation, or it may be a text which has been written by a

consortium of authors in a variety of languages, and then published in various language versions

Una vez aclarado el concepto en el que se basa el material utilizado en este estudio, conviene describir las características de este corpus:

- **Tipo de corpus.** Nos encontramos ante un corpus de textos paralelos bilingües español-árabe con una posible ampliación multilingüe.
- **Tipo de traducción.** Los textos de la L2 no son traducciones directas de la L1, ya que tanto la versión española como la versión árabe se han traducido a partir de un original, que en la mayoría de los casos está escrito en lengua inglesa.
- **Cobertura y representatividad.** Es un corpus que refleja el uso actual del lenguaje estándar en su variedad escrita tanto en el español como en el árabe.
- **Fuentes.** Todos los textos son documentos que proceden de instituciones internacionales pertenecientes a la Organización de las Naciones Unidas donde ambos idiomas, tanto el español como el árabe, tienen la calidad de lengua oficial. En su mayoría, los documentos son informes publicados por las siguientes instituciones: El Consejo de Seguridad, La Asamblea General, El Consejo Económico y Social, La Corte Internacional de Justicia, UNESCO: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura.
- **Propiedad intelectual.** El uso actual de los documentos no viola las leyes de propiedad intelectual, ya que se trata de fines académicos sin ánimo de lucro, según lo establecido por el principio de uso justo conocido en la propiedad intelectual.
- **Dimensión y tamaño.** La dimensión es bastante limitada; la versión actual del corpus contiene casi un millón de palabras en cada lengua (árabe: 901. 511 y español: 1.343. 225). No obstante, la muestra del corpus empleada aquí se reduce aproximadamente a 40.000 palabras en español más la parte correspondiente en árabe, tal y como se representa en la siguiente tabla:

	Español	Árabe
Número de *tokens*	39.496	26.179
Número de *types* (sin ruido textual[1])	3.986	4.796
Número de párrafos	664	660
Número de oraciones	1179	1173

- **Niveles de anotación del corpus.** El corpus está anotado a nivel estructural (párrafos, oraciones y *tokens),* y a nivel categorial. Para la anotación de las categorías gramaticales se han utilizado anotadores automáticos monolingües basados en reglas (Samy 2005). Además, el corpus está alineado en el nivel de las oraciones, es decir, que cada oración está relacionada con la(s) oración(es) en la traducción correspondiente.

[1] Por ruido textual nos referimos a problemas de procesamiento relacionados con la conversión de formatos como la presencia de doble espacios o caracteres no-reconocidos. Estos problemas han sido manejados para calcular el número preciso de los *types*, es decir, los tipos de palabras (en el sentido formal y técnico de secuencia de caracteres) en comparación con los *tokens* considerados como las ocurrencias individuales de cada uno de estos types.

2. Pragmatext, un modelo de anotación pragmática para corpus

2.1. Principios teóricos

A continuación presentamos muy resumidamente los cimientos teóricos en los que se sustenta nuestro modelo de anotación. Las fuentes de conocimiento en las que está basada esta concepción teórica pertenecen a los últimos descubrimientos en el estudio de la interacción mente-cuerpo por parte de las Ciencias Cognitivas (incluyendo la Lingüística Cognitiva), el conocimiento construido por la Pragmática Lingüística (sobre todo en lo referente a marcadores del discurso) y finalmente, la contemplación de dicho conocimiento a la luz de una perspectiva computacional. A continuación enumeramos brevemente los fundamentos teóricos del modelo de anotación:

1) La visión del ser humano de la realidad es subjetiva. Los procesos de percepción de la realidad se podrían definir como una interacción constante entre lo que está fuera y lo que está dentro de nuestra mente. Así, la mente intenta buscar una coherencia entre los datos que percibe y el conocimiento del mundo que tiene ya almacenado. Los procesos mentales son una constante reelaboración de información al servicio de la expectativa, la decisión y la planificación de la acción futura. Según estas razones de orden biológico, la verdad de los enunciados depende de la fuente de conocimiento en la que se apoye, del grado de compromiso que el hablante quiera adquirir con ella, del tipo de inferencias en las que se ha basado el hablante, etc.

2) Somos infodevoradores y la información que procesamos está regulada por el principio de relevancia: buscamos, de forma automática, mayores efectos contextuales con el menor coste de procesamiento.

3) El proceso de comunicación lingüística es inferencial, y también está regido, como cualquier otro proceso cognitivo, por el principio de relevancia, el cual juega un papel discriminador durante el proceso de interpretación-codificación lingüística de los enunciados por parte del hablante y de descodificación del mensaje e interpretación del mismo por parte del interlocutor. (Sperber & Wilson 1995)

4) A pesar de que los mensajes tienen un significado convencional, cuyo objeto de estudio pertenecería a la semántica, el significado último de un mensaje depende del contexto; de esta relación entre la interpretación y el uso del lenguaje en un contexto determinado se ocuparía la Pragmática.

5) Desde una perspectiva computacional, esta última sentencia dejaría en punto muerto a la Inteligencia artificial y a toda iniciativa que se planteara la comprensión y generación de lenguaje en términos automáticos, ya que al final el significado último de un enunciado depende de la interpretación que haga la mente de un individuo. Sin embargo, desde las Ciencias Cognitivas (pero también desde otros marcos de posicionamiento, como el de Goffman (1959) desde la Sociología) se está insistiendo en la importancia de los marcos a la hora de guiar la conducta y las inferencias mentales del sujeto en una interacción. Desde una perspectiva computacional, una manera de salir de este *cuello de botella* consiste en introducir conocimiento del mundo en los ordenadores (una de las apuestas que una vertiente de la Inteligencia Artificial hace por el diseño de sistemas expertos); otra posibilidad, dentro de la Lingüística Computacional, sería la de comenzar a explicitar conocimiento semántico y pragmático en los corpus a través de su anotación. En

cualquiera de las dos soluciones, debemos formalizar esta interacción entre lo cognitivo, lo cultural, lo social y lo lingüístico.

8) En nuestro modelo de anotación proponemos, por tanto, un camino diferente al que tradicionalmente la Lingüística Computacional ha seguido en lo que a etiquetado de corpus se refiere: dicho camino ha partido de las categorías gramaticales para acabar en el nivel semántico, y es, en buena parte, responsable de los grandes problemas de desambiguación que al final lingüistas e informáticos deben afrontar. En Pragmatext se parte, por el contrario, de la idea de que un fenómeno de carácter universal, por tener una razón cognitiva, se codifica lingüísticamente a través de diferentes niveles de la gramática y se materializa en formas lingüísticas concretas que cumplen esta función. Así por ejemplo, la evidencialidad es un fenómeno a través del cual se expresa la fuente de conocimientos en la que se apoya la verdad de un enunciado. La evidencialidad es un fenómeno universal porque es inherente a nuestra percepción, ya que no podemos saberlo todo de primera mano. Este fenómeno se expresa de diferentes formas según la lengua de la que tratemos, por ejemplo, para algunas lenguas del este, hay ciertos morfemas responsables de especificar esta información (Jakobson 1957). En español, algunas estrategias no están lexicalizadas, otras sí, como por ejemplo, los marcadores discursivos *por lo visto*, *relativamente*, *aparentemente*, etc.[2]

9) Los fenómenos de naturaleza semántico-pragmática que se marcan en Pragmatext son los que se citan a continuación: expresión de las emociones, relaciones argumentativas, modalización lingüística, evidencialidad, actos de habla y deixis social y textual. Estos fenómenos han sido estudiados por la Pragmática Lingüística, la Cortesía Lingüística, la Lógica, la Lingüística Textual, la Lingüística Cognitiva, el Análisis Crítico del Discurso, y el Análisis de la Conversación; y muchos de ellos se codifican a través de los marcadores discursivos.

2.2. Descripción del modelo

En este artículo vamos a aplicar este modelo con el fin de dar cuenta del significado de los marcadores discursivos encontrados en el corpus español, aunque, como ya hemos señalado, también estos fenómenos se expresan a través de otras estrategias gramaticales que derivan en otras formas lingüísticas. Por ejemplo, la atenuación discursiva también se puede expresar a través del uso de verbos modales conjugados en tiempo condicional: *Me gustaría hacerle una pregunta.* A continuación presentamos una tabla que ilustra esquemáticamente el modelo de anotación:

<table>
<tr><th>COGNICIÓN EN INTERACCIÓN</th><th colspan="2">PRAGMATICA</th><th colspan="2">LINGUISTICA</th></tr>
<tr><th>Estrategias</th><th>Fenómenos</th><th>Categorías</th><th>Categorías</th><th>Formas lingüísticas</th></tr>
<tr><td rowspan="2">¿Cómo razonamos?</td><td rowspan="2">Relaciones Argumentativas</td><td>Coargumentación</td><td rowspan="2">Conjunciones interjecciones Adverbios oracionales</td><td rowspan="2">además, pero, porque, en primer lugar, y, aunque</td></tr>
<tr><td>Contraargumentación</td></tr>
</table>

[2] Debemos a la Gramática Cognitiva esta visión de los fenómenos lingüísticos (Cuenca 1999).

		Concesión		Sintagmas Preposicionales Refranes …	...
		Reformulación …			
¿Cómo modalizamos la percepción cuando interactuamos?	Modalización discursiva	Intensificación			*seguramente, desde luego, claro*
		Atenuación			*Bueno*
		Interacción			*¿eh? ¿no?¿sabes?*
¿En qué fuentes apoyamos la verdad de nuestros enunciados?	Evidencialidad	El sentido de la vista de la persona			
		Otros sentidos			*Generalmente*
		Inferencia			
		Otra persona			*Por lo visto*
		Fuente escrita			*Según la Biblia*
		Fuente oral			*Quien a hierro mata, a hierro muere*
¿Cómo expresamos las emociones verbalmente?	Lenguaje de las emociones	Evaluación	Positiva		*Afortunadamente,*
			Negativa		*Desgraciadamente*
		Otros: Sorpresa			*Ah, qué fuerte*
¿Cómo conceptualizamos la percepción y cómo la convencionalizamos?	Metáfora	Dominio Origen : Cuerpo, Espacio, …		Colocaciones	*Tomar nota*
				Locuciones	*Echar de menos*
				Refranes	*Sarna con gusto no pica, pero mortifica*
		Dominio Meta: Tiempo, discurso, …			*a un tiro de piedra quien a hierro mata, a hierro muere*
¿Cuáles son nuestras intenciones?	Actos de habla[3]	Peticiones		Oraciones con entonación interrogativa	*Puedes cerrar la puerta?*
		Demandas de información …			*A qué hora sale el tren?*
¿Cómo nos referimos a otras partes del discurso y al interlocutor?	Deixis	Textual		Adverbios …	*anteriormente*
		Social		Vocativos	*Pibe, tío, macho*

[3] La tipología de actos de habla se centra en los enunciados interrogativos para un corpus restringido a un contexto determinado, con el que luego se entrena un sistema de diálogo hombre-máquina.

Antes de explicar cómo cobra forma este esquema de anotación de marcadores discursivos en lenguaje XML, nos gustaría primero exponer algunos problemas teóricos y computacionales de los marcadores discursivos, ya que nuestra formalización también está orientada a dar cuenta de ellos en la medida de lo posible.

2.3. Pragmatext y los marcadores discursivos: una perspectiva teórica y computacional

En este trabajo hemos intentado trazar un puente de comunicación entre la Pragmática Lingüística y la Pragmática Computacional en lo que al tratamiento de los marcadores discursivos se refiere. Buena parte de este compromiso pasa por adoptar una postura coherente con respecto a los siguientes hechos:

1. **Inventario de marcadores discursivos**: es momento de que la comunidad científica llegue a un acuerdo, por un lado, sobre qué es marcador discursivo y qué no, y por otro, sobre el número de marcadores discursivos existentes en una lengua así como de su distribución conversacional en términos diatópicos, diastráticos y diafásicos. Los corpus deben ayudarnos a esta tarea de recopilación.
2. **Ambigüedad categorial**: muchas palabras que juegan un papel discursivo tienen a su vez otras funciones en el nivel oracional, como nombres, adjetivos, etc. Ejemplos de marcadores ambiguos categorialmente son, por ejemplo, *bueno, entonces, primero, segundo, etc.*
3. **Polifuncionalidad**: Los marcadores discursivos están codificando información de diferentes clases (modal, ilocutiva, evidencial, deíctica, etc.) y algunos de ellos la expresan de forma simultánea. Además, marcadores como *pues, bueno, como*, etc. cumplen diferentes funciones discursivas en función de su posición y del tipo de discurso.
4. **Función discursiva y función oracional:** También debemos llegar a un acuerdo sobre si etiquetar las ocurrencias la conjunción *y* (conjunción con un valor coargumentativo), solo cuando une oraciones, o si debemos etiquetarla también cuando une otro tipo de cláusulas como, por ejemplo, sintagmas nominales.
5. **Idiomaticidad:** Necesitamos llegar a un acuerdo sobre cómo tratar la frecuente coaparición de más de un marcador discursivo en determinados contextos, tales como *pero si,* en oraciones del *tipo Pero si yo no he sido*, por ejemplo. Al tiempo que también debemos decidir si siguen siendo o no el mismo marcador *claro* y *claro que*; *o sea* y *o sea que*, etc.
6. **Localización automática en el texto:** Debemos desarrollar instrumentos de desambiguación tanto categorial como discursiva que tengan una cobertura y una precisión aceptables en la anotación automática de corpus.

Por otro lado, desde la perspectiva computacional, necesitamos sistematizar y formalizar los marcadores discursivos para poder operar con ellos, al tiempo que debemos considerar las limitaciones del trabajo con textos sin contexto (como son los corpus no-multimodales, como este) y que después se deben procesar automáticamente. Si implementáramos, por ejemplo, todos los valores que del marcador discursivo *bueno* ha establecido la bibliografía, tendríamos un etiquetado muy rico, pero debería ir acompañado de un etiquetado de corpus a nivel de enunciado que distinguiera la misma información;

pero todavía no existen corpus etiquetados a este nivel. Así pues, debemos acercarnos a unas definiciones básicas que se mantengan constantes si no en todos al menos en la mayoría de los contextos de realización.

2.4 Etiquetado en XML

A continuación presentamos el modo en que hemos formalizado el significado de los marcadores discursivos en lenguaje XML.

Una de las primeras decisiones que debemos tomar es si esta información debe ser declarada a nivel de atributos o a nivel estructural (esto es, a nivel de elementos) y por qué. Nosotros hemos decidido dejar el nivel de elementos para explicitar la estructura externa de los textos de la ONU: párrafos, enunciados y unidades con información pragmática, en este caso, los marcadores discursivos; mientras que hemos reservado el nivel de los atributos para explicitar información sobre estas partículas de orden discursivo.

Sobre un marcador del discurso se explicitará la siguiente información: (1) Un identificador, (2) Los lemas que compongan el marcador discursivo, (3) Categoría gramatical originaria, (4) Un rasgo que explicite si es una unidad fraseológica o no, y el tipo: colocación o locución, (5) Posición en la que aparece dentro de la *utterance* o enunciado: inicial, media o final, (6) Si es operador o conector, (7) Si posee contenido metafórico y, en el caso de que así sea, los campos semánticos de origen y destino implicados, (8) Si contiene contenido emocional, y en el caso de que así sea si es negativo, positivo o de otra naturaleza (sorpresa, por ejemplo), (9) Si expresa una relación argumentativa: generalización, focalización, hipótesis, etc., (10) Si es atenuante, intensificador o interactivo, (11) Si contiene un acto de habla, y por último, (12) Si posee un valor deíctico textual o social. A continuación presentamos un ejemplo de cómo se etiquetan en XML estas unidades:

(1)

<PI ID="1" Lema1="por" Lema2="ejemplo" GC="Prepositional_Phrase" DP="2" Range="operator" FU="Loc" MET="No" DR="Concretion" ED="No" MOD="No" EVI="No" SA="No" DEX="No">por ejemplo</PI>

En Pragmatext estos rasgos están separados y se pueden volver operativos o no, se pueden activar de una manera o de otra, dependiendo del marcador discursivo del que se trate. Esta es una de las ventajas del modelo de anotación, esto es, su flexibilidad a la hora de reflejar significados que remiten a fenómenos o dimensiones diferentes de la comunicación lingüística. . De esta manera, solucionamos a nivel teórico buena parte del problema de la ambigüedad de los marcadores discursivos, ya que cada uno de estos rasgos representa una dimensión de su posible contenido semántico.

3. Los marcadores discursivos en el corpus español de la ONU

Llegamos por tanto a la etapa del reconocimiento y etiquetado de los marcadores discursivos en un texto de habla oral. Para la resolución de esta tarea debemos plantearnos cómo vamos a gestionar el problema de la ambigüedad de los marcadores discursivos, una ambigüedad que como ya hemos señalado anteriormente, puede ser categorial pero también discursiva, en el caso de que no hayamos podido resolver este problema a nivel teórico.

Nosotras hemos manejado el problema de la ambigüedad en el corpus español de la siguiente forma. Hemos dividido los marcadores discursivos en cuatro grupos:

1. Marcadores no ambiguos. Ejemplo: *Generalmente*
2. Marcadores con ambigüedad categorial. Ejemplo: *Primero*
3. Marcadores con ambigüedad discursiva. Ejemplo: *Como*
4. Marcadores con ambigüedad categorial y discursiva. Ejemplo: *Bueno*

Los marcadores de tipo 1 se reconocen y se etiquetan de forma automática. Para desambiguar los marcadores de tipo 2, hemos recurrido a reglas de desambiguación contextual basadas en información prosódica, en posición discursiva y en información categorial. Para los marcadores de tipo 3, hemos seleccionado el significado más frecuente, y luego se ha verificado a mano. Para los marcadores de tipo 4 hemos aplicado las estrategias del tipo 2 y 3 simultáneamente, y luego se han revisado a mano los errores. Una vez que el corpus se ha etiquetado, se ha validado su buena formación como documento XML a través de una DTD.

3.1 Fenómenos pragmáticos y marcadores discursivos en el corpus de la ONU

De 1179 oraciones que forman la parte española del corpus, el anotador automático ha reconocido y etiquetado ocurrencias de uno o más marcador discursivo en 411 oraciones. El número total de ocurrencias de marcadores discursivos en español es 558 representando 83 tipos diferentes según la materialización lingüística del marcador discursivo. Sin embargo, si consideramos la función discursiva, las 558 ocurrencias representan 92 marcadores discursivos. Estos datos reflejan la ambigüedad discursiva, ya que una sola forma lingüística puede desempeñar diferentes funciones discursivas.

Por otra parte, es momento de aclarar que, tal y como ya se ha encargado de señalar frecuentemente la bibliografía, no todas las relaciones discursivas se explicitan a través de marcadores discursivos. En este sentido, nos gustaría destacar el estudio de Maite Taboada en el que se analizaron dos corpus, uno de conversaciones y otro de artículos de periódico, sobre los que intentó proyectar el modelo SDRT de Mann y Thompson. Esta autora concluyó que entre el 60% y el 70% de las relaciones discursivas no estaban expresadas explícitamente a partir de una forma lingüística (Taboada 2006).

A continuación se expondrán qué marcadores discursivos en el corpus de la ONU han dado forma a estas relaciones, y cuáles son sus ocurrencias en el corpus.

3.2 Las relaciones argumentativas

La tipología de relaciones entre enunciados que nos hemos planteado aquí se encuadra dentro de la preocupación por hallar una relación entre las formas lingüísticas y los modos de razonar y de expresar el pensamiento. Como sabemos, por un lado, tenemos a un sector de las Ciencias Cognitivas (Gardner 1987) con J. Fodor a la cabeza, que fundamentalmente defiende la existencia de un lenguaje propio del pensamiento, con el que el pensamiento opera, diferente a los sistemas lingüísticos. Durante mucho tiempo se ha aspirado a dar cuenta del lenguaje del pensamiento mediante el lenguaje de la lógica. Pero, sin embargo, hay bastantes investigaciones en el terreno de Teoría del razonamiento que confirman que la lógica del sentido común con la cual la gente se comporta en su vida

diaria, es diferente a la lógica que ha cultivado en la Filosofía (Lengrezi,1998). Por su parte, dentro del panorama lingüístico la Pragmática y de la Semántica, los lingüistas han reflexionado sobre algunas equivalencias entre conectivas lógicas y conectivas lingüísticas, pero la comparación no ha dado buen resultado (Escandell 1993). De nuevo, las conclusiones han sido parecidas, el contenido importa, determina el tipo de razonamiento. Anscombre y Ducrot (1994) defienden en su Teoría de la Argumentación algo parecido; estos autores señalan que los enunciados están orientados a defender unas posiciones argumentativas u otras. Dentro de esta argumentación que se lleva a cabo cuando se habla, los marcadores discursivos juegan un papel clave a la hora de guiar las inferencias, entendiendo por inferencia, el razonamiento que se expresa a través de varios enunciados.

Nosotras vamos a utilizar en nuestra tipología tipos de operaciones que tienen una larga tradición de análisis en los estudios descriptivos sobre marcadores del discurso. En ella hemos distinguido las siguientes clases:

1) Generalización y concreción.
2) Coargumentación, contraargumentación y concesividad
3) Hipótesis y condición
4) Reformulación
5) Síntesis
6) Topicalización
7) Causa, consecuencia, tiempo y finalidad

A continuación presentamos las tablas de frecuencia de marcadores discursivos asociados a la operación que están codificando.

Operación discursiva	**Frecuencia total en marcadores usados**	**Marcadores discursivos**	**Frecuencia del marcador**
Generalización	7	En general	7
Concreción	47	En particular	19
		Especialmente	8
		Como	6
		Particularmente	5
		Concretamente	4
		En concreto	1
		Tal como	2
		Por ejemplo	2
Coargumentación	224	Y	201
		E	20
		En principio	1
		Incluso	1
		En primer lugar	1
Coargumentación1	74	también	41
		asimismo	18
		Además	15
Coargumentación 2	2	Por otra parte	2
Coargumentación 3	5	Finalmente	3
		Por último	2

Contraargumentación	22	Pero	16
		Sin embargo	2
		No obstante	2
		Sino	1
		Mientras que	1
Hipótesis/condición	13	Una vez que	3
		Si	8
		Siempre y cuando	1
		Sin que	1
Causa	5	Porque	1
		Gracias a/al	2
		Debido a que	1
		Puesto que	1
Consecuencia	17	Pues	4
		Por consiguiente	3
		En consecuencia	3
		Como resultado de	2
		De resultas de	1
		A consecuencias de ello	1
		Ya que	1
		Por ello	1
		Por cuanto	1
Concesión	13	Aunque	13
Reformulación	5	A saber	4
		Es decir	1
Topicalización	37	Respecto de	13
		Con respecto a/al	8
		En cuanto a/al	4
		A este respecto	3
		A tal respecto	2
		Respecto del	1
		Desde el punto de vista x	2
		En relación con	1
		En este sentido	2
		En este contexto	1
Síntesis	1	Principalmente	1
Tiempo	15	Cuando	7
		Al mismo tiempo	1
		Al tiempo que	2
		Mientras	1
		Tan pronto como	1
		Desde que	1
		A la vez que	1
		Entre tanto	1
Finalidad	32	Para que	29
		A fin de que	2

		A tal fin	1

Otras operaciones discursivas son la **digresión** y la **planificación del habla** (apoyos vocálicos) pero no las vamos a tratar aquí por no aparecer en la ONU.

3.3 La modalización discursiva

La modalización discursiva ha sido objeto de estudio por parte de diferentes ramas del conocimiento, desde la Filosofía hasta la Pragmática. Nosotros entendemos la modalización como un proceso cognitivo. Cada ser humano posee un modelo de mundo. La modalización que se lleva a cabo durante el discurso puede estar motivada por la cognición y regulada por unas normas que conducen no solo a la comunicación sino al refuerzo de lazos de solidaridad, ya que también somos seres sociales. Por ello, el ser humano puede atenuar la verdad de sus enunciados bien porque no esté seguro de dicho valor de verdad, bien porque intenta buscar acuerdo con su interlocutor. De la misma forma, la intensificación del discurso puede estar causada por un compromiso total con la verdad de los enunciados que se está verbalizando, o puede estar causada por la intención de intensificar el acuerdo y por tanto las relaciones sociales con el interlocutor.

Por otra parte, dentro de las operaciones de modalización discursiva nos gustaría diferenciar un tipo más, el tercero y el último: la interacción. Durante el proceso de interacción, hay formas lingüísticas que cumplen la función de buscar o bien acuerdo o bien necesidad de saber si el mensaje está siendo interpretado en el sentido en que al hablante le gustaría; ejemplos de estos marcadores son *¿eh? ¿no? ¿entiendes?¿sabes? ¿me explico o no me explico?*.

Modalización	Intensificación	2	Verdaderamente	1
	Atenuación		Posiblemente	1

3.4 La evidencialidad

Como ya se ha explicado, a través de estos marcadores se marca lingüísticamente la fuente del conocimiento.

Como sabemos hay una relación entre evidencialidad y la modalidad. La expresión de evidencialidad tiene un valor argumentativo, una autoridad en el juicio sobre el valor de verdad del enunciado. Así por ejemplo, un marcador como *por lo visto* tendría un contenido evidencial pero también un contenido atenuante, ya que con dicho marcador el hablante se distancia del valor de verdad de su enunciado, atenuando la verdad del mismo.

Nosotras hemos diferenciado las siguientes fuentes de conocimiento: (1) vista (2) otros sentidos (3) inferencia lógica (4) testimonio de otro (5) fuente oral (6) fuente escrita. En el futuro, la Antropología y la Sociología deben colaborar con la Lingüística para jerarquizar las fuentes en función de la autoridad que cada una de ellas tiene en cada cultura. Nosotros no hemos establecido una jerarquía entre ellas. Solamente etiquetamos las expresiones que aluden a estos tipos. Los marcadores de evidencialidad en el corpus de la ONU son:

Evidencialidad	4	aparentemente	1

		Relativamente	1
		A juicio de ellos	1
		En su opinión	1

3.5. Deixis: El término de deixis se utiliza para denominar a las formas lingüísticas cuya referencia depende del contexto. En este sentido, hemos diferenciado dos tipos de maracadores deícticos: aquellos que hacen referencia a partes del texto y aquellos que hacen referencia al interlocutor. Aunque en este corpus solo hayamos encontrado el primer tipo.

Deixis	textual	De esta forma	1

4. Los marcadores discursivos del corpus árabe de la ONU

Para la anotación de los marcadores discursivos en árabe, se ha desarrollado un módulo de procesamiento que tiene como entrada tres fuentes de información:

1. La información de los marcadores discursivos anotados en el corpus español.
2. Los datos de alineamiento y la anotación de categorías gramaticales.
3. Un lexicón bilingüe de marcadores discursivos español-árabe creado automáticamente a partir de los marcadores del español.

Para la última fuente de información, la lista de marcadores discursivos españoles se ha traducido automáticamente a través de un sistema de traducción automática y un diccionario electrónico. Con las traducciones proporcionadas de ambas fuentes se ha creado el lexicón bilingüe. Para la traducción automática, se ha utilizado el sistema *Google Translate* disponible en línea por Internet. Para el diccionario bilingüe, se ha utilizado el diccionario BetaWikiled Online Dictionary (Spanish-Arabic)[4]. Cabe señalar que el *Google Translate* no ofrece la opción de español-árabe, por lo tanto la traducción se ha llevado de forma indirecta a través del inglés en dos pasos: español-inglés e inglés-árabe.

Hemos recurrido a esta estrategia para maximizar el uso de las fuentes disponibles y automatizar en la medida de lo posible el proceso de anotación del corpus árabe. De este modo, se introduce una estrategia eficaz y se ahorra el tiempo y el esfuerzo manual requerido en el caso de la anotación manual. Sin embargo, hay que admitir que esta estrategia podría tener sus desventajas en cuanto al reconocimiento de marcadores discursivos propios del corpus árabe que no aparecen en el corpus español (considerado como el punto de partida). A pesar de ello, una comparación eficaz entre las ventajas logradas y las desventajas demuestra la eficiencia de esta aproximación en términos de tiempo y coste, dado que es una técnica reconocida en el procesamiento del lenguaje natural para acelerar el proceso de creación de recursos que suele caracterizar por su alto coste.

Por último y además de las fuentes arriba mencionadas, se utilizan algunas heurísticas sobre la posición de la ocurrencia del marcador discursivo y los signos de puntuación que

[4] http://www.wikiled.com/spanish-arabic-Default.aspx

aparecen con los marcadores. Se recurre a estas heurísticas en el caso de que no se pueda localizar un candidato con la ayuda de las fuentes.

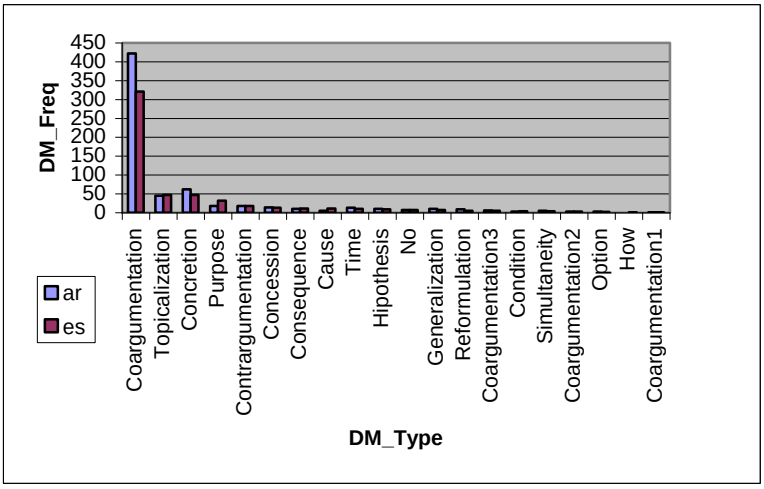
El procedimiento adoptado consiste en lo siguiente: primero, para cada oración en el corpus español, se extraen los marcadores discursivos anotados con los atributos asignados. Segundo, para cada marcador en la oración, se busca en el lexicón bilingüe las posibles traducciones en árabe. Tercero, una vez extraídas las traducciones en árabe y con la ayuda de la información proporcionada por el alineamiento se busca alguna ocurrencia de estas traducciones en la(s) oración(es) alineadas del corpus árabe. Cuarto, si se localiza una ocurrencia, se etiqueta con los mismos atributos de su correspondiente español. Quinto, si no se localiza ningún candidato, se recurre a las heurísticas sobre la posición, por ejemplo, si el marcador ocurre en una posición inicial o intermedia delimitado con los signos de puntuación. Sexto, para los atributos de información sobre categorías gramaticales, se recurre a la información categorial proporcionada en el corpus y se asignan las categorías a los atributos correspondientes. Uno de los desafíos en este respecto ha consistido afrontar el problema de la elevada frecuencia de los clíticos en el árabe, dado que un mismo token puede que esté formado por uno o más unidades gramaticales. Un caso parecido en español es la contracción de la preposición y el artículo como en "del" donde en un mismo token existen dos categorías gramaticales. En árabe este fenómeno es muy frecuente y las unidades gramaticales pueden llegar hasta cuatro como en el ejemplo siguiente:

preposición+artículo+nombre	**بالتحديد (en particular)**

La salida de este módulo consiste en el corpus árabe con los marcadores discursivos anotados con la misma información de los atributos de sus correspondientes en español.

5. Evaluación de los resultados

Los resultados de la anotación del corpus árabe a partir de la anotación del español demuestran que en muchos casos el módulo automático ha acertado con un porcentaje de 80,4%, ya que ha detectado correctamente 449 marcadores en árabe de los 558 marcadores en español.



Lo que nos interesa aquí desde una perspectiva lingüística es estudiar las estrategias por las cuales se han traducido los marcadores discursivos. Analizar los resultados ha revelado una serie de observaciones. Estas observaciones se pueden considerar como motivos de errores y ambigüedades que ha afectado el procesamiento automático, pero que a la vez revelan fenómenos propios del proceso de la traducción y la formulación de los procesos pragmáticos en cada idioma. Estas observaciones se mencionan a continuación.

Para empezar, un modelo automático parte de la hipótesis de que la traducción de marcadores discursivos es una relación de uno-a-uno. Esta premisa no se cumple en la realidad. En muchos casos, el traductor opta por la omisión de un marcador discursivo o recurre a otras estrategias lingüísticas en la lengua destino. Este fenómeno explica el porcentaje de errores en la detección automática.

En muchos casos, la función discursiva se materializa en español a través de un marcador discursivo, mientras que en árabe la misma función discursiva se materializa en paráfrasis verbales. Es decir, cada lengua adopta estrategias diferentes para reflejar funciones discursivas como la consecuencia, la co-argumentación, la hipótesis, etc. Un ejemplo bastante frecuente es la traducción de los marcadores de finalidad como *para que*, *con el fin de*, (en ejemplos como "hacer una llamada internacional para/con el fin de") a una paráfrasis verbal de tipo " طالب ب" "exhortar a/llamar a".

En algunos casos, se ha observado que en la traducción a la lengua2, se puede optar por una operación discursiva diferente a la utilizada en la lengua1. Los casos de este tipo no son frecuentes, pero sí se dan, ya que en algunos casos, el traductor opta por un marcador de concreción en vez de un marcador de contra-argumentación.

El marcador discursivo de co-argumentación por excelencia "y", en árabe "و" ocurre en árabe con mucha frecuencia y es común utilizarlo con otros marcadores. Este marcador ha causado bastante ruido en el procesamiento, ya que presenta una ambigüedad de carácter sintagmático. Además, según las convenciones ortográficas del árabe moderno la "و" ocurre en el mismo token al que precede sin utilizar espacios para delimitarlo. Esto causa mucha ambigüedad formal porque cuando ocurre en una posición inicial de un token, no se puede distinguir si se trata de un carácter inicial que forma parte de la palabra o si se trata de una conjunción o un marcador discursivo.

Por otro lado en el caso de las heurísticas, aunque, a primera vista, se puede decir que normalmente los marcadores discursivos ocurren en posiciones parecidas y delimitados con los mismos signos de puntuación, en el análisis real de los resultados de este módulo, se ha observado que, en muchos casos, las heurísticas han dado mejores resultados en las posiciones iniciales de ocurrencia, mientras que en los casos de posiciones intermedios, las heurísticas han fallado en varias ocasiones por varios motivos como el cambio de la posición en cuanto al orden del segmento, la omisión o el uso de delimitadores diferentes. Asimismo, cabe señalar que el uso de los signos de puntuación en árabe es una práctica bastante reciente en la ortografía árabe. Por eso, su uso no sigue unas reglas bien definidas y se omiten en carios casos.

Por último, hay que tener en cuenta que si se estudia el corpus árabe desde una perspectiva monolingüe, es muy probable que se detectaran casos en que en el texto traducido aparece un marcador discursivo propio del texto árabe sin que sea utilizado en el texto español. Sin embargo, detectar estos casos requiere un procesamiento monolingüe que no lo hemos seguido en este experimento, ya que en el presente estudio, partimos del corpus español para localizar equivalentes en el corpus árabe.

Tras estas observaciones, señalamos algunos de los resultados de los marcadores discursivos detectados en árabe.

Marcadores de finalidad

Marcador Discursivo	Frecuencia
من أجل	7
ووصولا إلى هذه الغاية،	1
بحيث	1
كي	1
لكي	1

Marcadores de contra-argumentación

Marcador	Frecuencia
لكن	8
بل	4
إلا	3
غير	1
ومع هذا،	1

Marcadores de condición

Marcador	Frecuencia
بمجرد	1
فور توجهها إلى العراق ،	1
بعد	1

Marcadores de Topicalización

Marcador	Frecuencia
بشأن	15
من حيث	2
وفي هذا الصدد،	2
في جملة أمور،	1
في هذا الصدد	1
في هذا الشأن	1
لا سيما	1
إلا أن	1
و	1
وفيما يتعلق	1
متعلق ب	1

Marcadores de co-argumentación

Marcador	Frecuencia
و	44
أيضا	24
كما	12
فضلا عن	6
كذلك	4
وعلاوة على ذلك،	2
وقد أيد بعض أعضاء المجلس **كذلك** وضع ولاية أكثر قوة،	2
إضافة إلى	2
لا سيما	1
وبالإضافة إلى ذلك،	1

Finalmente, la salida final del corpus paralelo anotado con los marcadores discursivos se representa en la siguiente forma donde se resaltan los marcadores discursivos con efectos visuales utilizando los colores.

1 El Secretario General Adjunto se refirió **en particular** a los progresos logrados en relación con la iniciativa de paz de Djibouti , **así como** a la situación política y humanitaria en Somalia .	1 وأشار وكيل الأمين العام ، بوجه خاص، إلى التقدم الذي أحرز بشأن مبادرة جيبوتي للسلام، بالإضافة إلى الحالة السياسية والإنسانية في الصومال .
2 **Con respecto al** plan de paz de Djibouti , informó a los miembros del Consejo de las actividades del Presidente de Djibouti , Omar Guelleh , en los países de la región ; indicó **también** que la iniciativa había sido bien recibida por la sociedad somalí .	2 وقد أبلغ أعضاء المجلس عن الخطوات التي اتخذت من جانب رئيس جيبوتي في بلدان المنطقة فيما يتعلق بخطة جيبوتي للسلام .
2 **Con respecto al** plan de paz de Djibouti , informó a los miembros del Consejo de las actividades del Presidente de Djibouti , Omar Guelleh , en los países de la región ; indicó **también** que la iniciativa había sido bien recibida por la sociedad somalí .	3 وأوضح أيضا أن المبادرة استقبلت استقبالا طيبا من قبل المجتمع الصومالي .
3 **En cuanto a** la situación política y militar , dijo que durante enero y febrero se habían comunicado incidentes de bandidismo , **así como** confrontaciones entre los clanes .	4 أما فيما يتعلق بالحالة السياسية والعسكرية، فقد ذكر أنه قد ترددت أنباء عن وقوع حوادث لقطع الطرق، بالإضافة إلى مواجهات بين العشائر خلال شهري كانون الثاني يناير و شباط فبراير .

5. Conclusiones y trabajo futuro

Hemos presentado en este artículo una aproximación al campo de los marcadores discursivos desde una perspectiva computacional. Hemos apostado por un modelo de anotación pragmática basado en fenómenos de naturaleza cognitiva y social, y que remiten al conocimiento generado en la Pragmática de corte teórico. La motivación principal es, por una parte, sistematizar este conocimiento y, por otra crear, un estándar sobre anotación pragmática de corpus basado en rasgos que aluden a clasificaciones que se pueden activar forma simultánea. De esta manera, se propone una solución computacional a algunos de los problemas teóricos de estas partículas. Los fenómenos a los que se refieren la semántica de los marcadores son de origen cognitivo e influyen en la verdad de los enunciados. Con esto pretendemos dar un toque de atención a algunos planteamientos computacionales que se basan en codificación binaria Verdadero/Falso de los enunciados. Como vemos, las lenguas naturales codifican por el contrario la relatividad de la verdad de los mismos.

Hasta aquí en lo que respecta a la parte teórico-descriptiva del artículo. En cuanto al lado computacional, hemos intentado acercarnos al reconocimiento y clasificación automática de marcadores discursivos en dos lenguas, con el fin de estudiar la traducción de estas partículas en el corpus paralelo español-árabe de la ONU. Las estrategias computacionales seguidas para cumplir dicho objetivo han tenido diferente grado de éxito y en el futuro se apostará por la aplicación de dichas estrategias a corpus de otros géneros discursivos y con mayor número de palabras. El resultado final de la investigación ha sido la obtención de un corpus paralelo en formato XML con los enunciados alineados y con los marcadores discursivos etiquetados y clasificados según su valor en el discurso. Las aplicaciones de este corpus pueden ser de naturaleza muy diversa. Nosotras hemos desarrollado un interfaz Web de consulta donde el usuario podrá extraer información sobre cómo aparece traducido un marcador discursivo en particular en el corpus que se elija. No obstante, existen otras aplicaciones, como la enseñanza de lenguas o la inducción automática de estructura retórica con diferentes fines, como por ejemplo, la generación automática de resúmenes.

6. Bibliografía

Alonso, Laura (2002), Irene Castellón y Lluís Padró. "Lexicón Computacional de Marcadores del Discurso", *Procesamiento del lenguaje natural*, 29, 2002, pp 239-246.

Anscombre, Jean-Claude y Oswald Ducrot (1994). *La argumentación en la lengua*, Madrid: Gredos.

Casado, Manuel (1998): "La gramática del texto y los marcadores del discurso", en Martín Zorraquino, Antonia y Montolío, Estrella (coords.) (1998). Los *marcadores del discurso. Teoría y análisis*, Arco Libros. Madrid .

Cuenca, Mª Josep, Joseph Hilferty (1999). *Introducción a la Lingüística Cognitiva*, Barcelona: Ariel.

Escandell, Victoria. (1993). *Introducción a la pragmática*. Barcelona: Ariel.

Fernández, Patricia (2006). "Hacia una propuesta de clasificación de unidades de traducción". *RAEL: revista electrónica de lingüística aplicada*, ISSN 1885-9089, Nº. 5, pags. 141-154.

Gardner, Howard (1987). *La nueva ciencia de la mente: Historia de la psicología cognitiva*. Barcelona: Paidós.

Goffman Erving (1959). *La presentación de la persona en la vida cotidiana*. Buenos Aires-Madrid: Amorrortu-Murguía, 1987

González-Ledesma, Ana (2007). "Pragmatext, Annotating the Spanish C-ORAL-ROM Corpus with Pragmatic Knowledge" , en *Proceedings of 4th Corpus Linguistics Conference, University of Birmingham*, 27-30 July.

Granger, Sylvyanne (2003). "The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies?". En *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, Granger, Sylviane, Jacques Lerot y Stephanie Petch-tyson (eds.) Amsterdam/New York, NY, 2003.

Hackey, Leo ed. (1998): *Pragmatics of translations* Clevedon [England] ; Philadelphia, Multilingual Matters.

Jakobson, Roman (1957). "Shifters, verbal categories and the Russian verb". En *Selected Writings* Vol. 2 Cambridge University (1971). The Hague: Mouton, 130-147

Mann, William y Thompson, Sandra (1988). "Rhetorical Structure Theory: A theory of text organization". En L. Polanyi (ed.) *The Structure of Discourse*, Ablex, 1988.

Marcu, Daniel (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, November 2000. ISBN 0-262-13372-5.

Portolés, José (2004). *Pragmática para hispanistas*. Madrid: Síntesis.

Prada, José y Guillermo Moncecchi (2003). "Reconocimiento eficiente de marcadores del discurso en español". VIII Simposio Internacional de Comunicación Social, Santiago de Cuba, Cuba, January, 2003

Samy, Doaa (2005): *Recursos bilingües de ingeniería lingüística para el procesamiento del español y el árabe*. Tesis doctoral. Universidad Autónoma de Madrid.

Samy, Doaa, Moreno-Sandoval, Antonio, Guirao, José M. y Alfonseca, Enrique (2006): Building a Multilingual Parallel Corpus Arabic-Spanish-English. In *Proceedings of International Conference on Language Resources and Evaluation LREC-06*, Genoa, Italy.

Somers, H. (2001). Bilingual Parallel Corpora and Language Engineering. En *Proceedings of Anglo Indian Workshop "Language Engineering for South Asian Languages" LESAL*, Mumbai. Disponible en: http://www.emille.lancs.ac.uk/lesal/somers.pdf

Sperber, Dan and Wilson Deirdre (1995). *Relevance: Communication and cognition* (2nd ed.) Oxford: Blackwell, 1995.

Taboada, Maite (2006). "Discourse markers as signals (or not) of rhetorical relations", *Journal of Pragmatics*, Volume 38, Issue 4, April 2006, Pages 567-592, Focus-on Issue: The Pragmatics of Discourse Management