CORRELACIONES EN EUSKERA ENTRE LAS RELACIONES RETÓRICAS Y LOS MARCADORES DEL DISCURSO[1]

MIKEL IRUSKIETA QUINTIAN, ARANTZA DIAZ DE ILARRAZA SANCHEZ Y MIKEL LERSUNDI AYESTARAN
Grupo IXA de procesamiento del Lenguaje Natural, Universidad del País Vasco

1. INTRODUCCIÓN

En este artículo analizamos la distribución e interrelación de las unidades discursivas que componen un discurso. Presentamos los resultados de un estudio empírico sobre las correspondencias entre las relaciones de coherencia, sobre su señalización morfosintáctica, sobre un estudio comparativo del orden canónico del inglés y del euskera, sobre la posición discursiva del marcador del discurso y de la ambigüedad de los marcadores del discurso. En nuestra opinión, la Teoría de la Estructura Retórica (*RST*) de Mann y Thompson (1987) proporciona una base teórica adecuada, ya que: (a) es una teoría general para cualquier lenguaje y (b) es una teoría que ha sido aplicada a análisis computacionales (Marcu 2000); además se ha desarrollado una aplicación *RSTTOOL* (O'Donnell 1994), para el etiquetado de textos basado en esa teoría.

Uno de los mayores inconvenientes para el estudio computacional de las relaciones de coherencia es su escasa señalización morfosintáctica. Para Knott y Dale (1994: 10) no es posible hablar de relaciones sino existen marcadores del discurso. En cambio para Taboada (2006) esta afirmación tiene asociada una incógnita, ¿a qué es debido que una gran parte de las relaciones retóricas no estén señaladas?

1 Este artículo ha recibido ayuda del proyecto HUM2007-65966.CO2-022

Partiendo de una u otra de las posiciones el estudio de las correlaciones entre las relaciones retóricas y los marcadores del discurso es radicalmente distinto.

En Iruskieta *et al.* (2008) se enfocó el problema partiendo de los marcadores del discurso. Y en este artículo se enfoca el problema desde las relaciones retóricas. El artículo está organizado de la siguiente manera: en la segunda sección, presentamos el corpus utilizado; en la tercera sección se explica la correspondencia entre las relaciones retóricas y los marcadores del discurso, y se exponen los cinco aspectos objeto de estudio: (a) la ayuda del análisis morfosintáctico para la obtención de los marcadores del discurso; (b) la correspondencia entre marcadores de discurso y relaciones retóricas; (c) el orden canónico de las relaciones retóricas; (d) la posición de los marcadores del discurso; (e) la ambigüedad para el caso de que un mismo marcador de discurso pueda señalar más de una relación retórica. Por último, en la sección cuarta presentaremos las conclusiones y trazaremos el trabajo futuro.

2. LA TEORIA Y EL CORPUS

La búsqueda de la coherencia en un texto es exitosa en el caso de que no existan secuencias ilógicas. En textos bien argumentados a cada parte del texto se le asigna una función; gracias a las relaciones retóricas es posible establecer una jerarquía entre todas estas partes del texto. Mann y Thompson (1985: 258) introducen secuencias ilógicas para demostrar el espíritu de la incoherencia y así poder saber qué es la coherencia. Nosotros nos vamos a concentrar en analizar las relaciones entre las partes que componen el discurso.

Para realizar este estudio hemos elegido al azar 10 textos del Corpus de Referencia para el Procesamiento del Euskera (*EPEC*) (Aduritz *et al.* 2006). Los textos son periodísticos y corresponden a noticias acerca de diferentes temas: política (4 textos), deporte (3 textos) y otros (3 textos); en total 1442 palabras. *EPEC* se ha construido para el desarrollo de aplicaciones y sistemas automáticos del lenguaje vasco. Está etiquetado a distintos niveles lingüísticos: nivel morfosintáctico (Aldezabal *et al.* 2007a), nivel sintáctico superficial (Aldezabal *et al.* 2007b) y nivel semántico (Agirre *et al.* 2005). Nuestro objetivo es el de poner los fundamentos para añadir la información del nivel discursivo. El etiquetado de las relaciones retóricas ha sido realizado con la clasificación extendida de las relaciones retóricas explicada en Mann y Taboada (2008). Esta clasificación tiene 33 relaciones. Para la anotación se ha utilizado la aplicación *RSTTOOL* de O'Donnell (1994, 2000).

3. RELACIONES RETÓRICAS Y MARCADORES DEL DISCURSO

Llamamos marcador de discurso a los elementos morfosintácticos que señalan relaciones entre las unidades del discurso. Los marcadores del discurso pueden ser conjunciones coordinativas, conectores, conjunciones subordinantes, locuciones y modificadores oracionales.

Hemos analizado un total de 149 relaciones retóricas, de ellas 62 estaban señaladas con marcadores del discurso (41,6 %). Aunque el tamaño del corpus analizado es

pequeño, con esos datos observamos que la clasificación extendida de la *RST* nos aporta una base sólida para nuestro objetivo. Además, la herramienta *RSTTOOL* es de gran utilidad para el etiquetado del corpus *EPEC*. En nuestro estudio hemos observado que el 41,6 % de las relaciones retóricas aparecen señaladas con marcadores del discurso. Ese es similar al que da a conocer Taboada (2006: 579) para los textos periodísticos del ingles extraídos del *Wall Street Journal*: 43,48 %.

3.1. Obtención de las relaciones retóricas y los marcadores del discurso

En este apartado presentamos los pasos segundos para determinar la influencia de la información morfosintáctica en la detección de los marcadores de discurso. La metodología utilizada ha sido la siguiente: (a) Anotación de las relaciones retóricas; (b) Detección de los marcadores del discurso; (c) Identificación de correspondencias entre relaciones y marcadores.

La búsqueda de candidatos se ha realizado en base a la información lingüística contenida en *EPEC* partiendo de la información morfosintáctica, sintáctica y una combinación de ambas.

En la tabla 1 se indican los resultados obtenidos en este proceso. Los resultados se han comprobado manualmente, determinando de esa forma los errores y las carencias de la búsqueda. En la 1ª columna indicamos el tipo de información utilizada para la búsqueda de los marcadores; en la 2ª columna el número de candidatos obtenidos de manera automática; en la 3ª los errores o falsos candidatos obtenidos de cada que cada fuente lingüística; en la 4ª columna las carencias o el número de candidatos que el proceso automático no ha podido detectar.

	Automático	**Manual**	
Información lingüística	**Candidatos para MD**	**Falsos MD**	**Cantidad de MD no identificados**
Morfosintaxis	112	48	22
Sintaxis	118	44	12
Morf + Sint	142	62	6

Tabla 1. Búsqueda de candidatos.

Aunque se hayan recogido muchos falsos candidatos la búsqueda más provechosa ha resultado el de la suma de las dos fuentes; ya que, por una parte, es la que más candidatos ha encontrado (sólo faltan pronombres numerales y locuciones) y, por otra parte, podría implementarse una sencilla gramática para la eliminación de estos falsos candidatos.

3.2. Correspondencia entre las relaciones y los marcadores

La tabla 1 muestra información sobre la correspondencia entre los marcadores de discurso y las relaciones retóricas para aquellas relaciones que presentan una frecuencia de aparición superior a cinco. En la 1ª columna están las relaciones retóricas encontradas en los textos. En la 2ª columna se indica el número de relaciones retóricas identificadas. Finalmente, en la 3ª columna hemos puesto el número de marcadores del discurso que señalan las relaciones retóricas.

Relaciones retóricas	Cantidad	MD
Fondo rst	28	5
Elaboración rst	23	4
Conjunción multinuc	19	13
Preparación rst	17	0
Circunstancia rst	10	4
Secuencia multinuc	10	9
Concesión rst	8	8
Contraste multinuc	5	5
TOTAL	**120**	**48**

Tabla 2. Correlación cuantitativa sobre las relaciones retóricas y los marcadores del discurso.

A partir de los datos de la tabla 2 observamos 3 conductas posibles en la correspondencia entre las relaciones retóricas y los marcadores del discurso:

- Las relaciones que están señaladas siempre por marcadores del discurso. En este grupo encontramos las relaciones de *concesión* y de *contraste*.
- Las que están señaladas algunas veces, por ejemplo las relaciones de *fondo*, de *circunstancia*, de *conjunción* y de *elaboración*.
- Las que no están señaladas. En este caso solo hemos dado con una relación, la relación retórica de *preparación*.

3.3. Orden canónico de las relaciones

El orden en el que se presentan las unidades en una relación retórica puede ser considerado como dato a la hora de identificar el tipo de la relación retórica. En las relaciones retóricas existen dos tipos de orden: (1) La unidad satélite precede al núcleo: satélite-núcleo, y (2) El núcleo precede a la unidad satélite: núcleo-satélite.

En nuestra observación existe una pequeña diferencia en el orden canónico de las relaciones nucleares de las que Mann y Thompson (1987: 17) propusieron. Sólo hemos considerado, como en el caso anterior, aquellas relaciones que tienen más de 5 instancias. En la 1ª columna de la tabla 3 están las relaciones retóricas, en la 2ª los datos sobre el tipo de orden satélite-núcleo y en la 3ª el tipo de orden núcleo-satélite.

Relaciones retóricas	Satélite-núcleo	Núcleo-satélite
Fondo rst	1	27
Circunstancia rst	5	5
Concesión rst	3	5
Elaboración rst	-	23
Preparación rst	17	-
Total	**17**	**60**

Tabla 3. Orden genérico de las relaciones retóricas.

El tipo de orden de la *preparación* es por definición satélite-núcleo, por lo tanto no puede haber diferencias en lengua alguna. En cambio los datos son escasos e imprecisos para determinar el orden de la relación de *circunstancia*. Sucede lo mismo con los datos sobre el orden de la *concesión*.

Si comparamos con lo presentado para el inglés (Mann y Thompson 1987) podemos determinar y comparar el orden de las otras dos relaciones observando una igualdad y una diferencia con nuestro estudio. Igualdad: la *elaboración* tiene el mismo orden canónico en inglés y en euskera. Diferencia: el *fondo* tiene un orden diferente en euskera con respecto al inglés, ya que la unidad satélite suele estar a la izquierda.

3.4. La posición de los marcadores

Cuando hablamos sobre la posición de los marcadores del discurso podemos hablar de su posición en términos sintácticos o en términos discursivos. Lo hacemos en términos discursivos en la Tabla 4. En la 1ª columna se especifica el tipo de relación retórica: (a) por un lado, la relación nuclear que está compuesta de un núcleo y de una unidad satélite, es la unidad satélite la que marcara la relación retórica que posee con el núcleo; y (b) por otro lado, la relación multinuclear está compuesta con núcleos relacionados entre sí. En la 2ª columna el tipo de la unidad satélite o núcleo, en el caso de la relación multinuclear, como ya hemos mencionado, solo puede haber relación del mismo nivel entre los núcleos. Considerando que tanto el núcleo como la unidad satélite pueden encontrarse en cualquier orden entre sí, hemos indicado todas las combinaciones posibles de la posición del marcador del discurso en la 3ª columna. Y en la 4ª la cantidad de instancias en cada posición.

Tipo de relación	Tipo de unidad	Posición del MD	Cantidad
Relación nuclear	Núcleo	1ª unidad	1
		2ª unidad	3
	Satélite	1ª unidad	7
		2ª unidad	21
Relación multinuclear	Más de un núcleo	1ª unidad	0
		2ª unidad	28
		En todas las unidades	2
		TOTAL	**62**

Tabla 4. Posición discursiva de los marcadores del discurso.

La posición de los marcadores del discurso varía tratándose de una relación nuclear o de una relación multinuclear.

a. En las relaciones nucleares sólo en un 25,8 % de los casos el marcador del discurso esta en la 1ª unidad, siempre que no se distingan entre satélite o núcleo. Por lo tanto, en un 74,2 % de los casos ha sido localizado en la 2ª unidad. Los marcadores del discurso que han aparecido en la 1ª unidad son conjunciones subordinantes; estas conjunciones, a diferencia de otros marcadores, marcan la misma relación y el mismo significado estando en una u otra posición. Con los conectores, las conjunciones, las locuciones y los modificadores oracionales el cambio de posición puede cambiar el tipo de relación y el significado.
b. En las unidades multinucleares no ha habido ningún caso en el que el marcador del discurso esté únicamente en la 1ª unidad. En un 93,3 % de los casos han sido localizados en la 2ª unidad. Y el resto de los casos (6,7 %) se debe a que la coordinación distributiva se realiza distribuyendo marcas morfosintácticas en cada unidad discursiva.

3.5. La ambigüedad

Si algún marcador del discurso en contextos diferentes señala más de una relación retórica, entonces tomamos ese marcador del discurso como inespecífico o ambiguo, aún cuando en cada contexto en que se ha encontrado no sea ambiguo para el analista, sí lo es para el análisis computacional. Si entre dos unidades cabría interpretar más de una relación retórica y esa ambigüedad la crease el marcador del discurso hablaríamos de una ambigüedad real. Por lo tanto cuando en este artículo hablamos de la ambigüedad hablamos de una ambigüedad abstracta que proviene del análisis y no de una ambigüedad que existe en el texto. Marcu (2000: 94) señala 3 ambigüedades: a) el conector tiene dos funciones, o bien oracional o bien discursivo; b) el conector no determina la longitud de

sus unidades discursivas; y c) un conector puede señalar más de una relación retórica. Cuando en este trabajo hablamos de ambigüedad nos referimos a esa tercera ambigüedad. Marcu 2000 (81-82) señala otro fenómeno, a nuestro entender ligado a la ambigüedad: las relaciones múltiples. En el trabajo realizado no se ha considerado que pudiese haber tales relaciones múltiples.

En un trabajo anterior, Iruskieta *et al*. (2008: 4) se trató la ambigüedad desde otro método analítico. En este trabajo se han encontrado además de los dos marcadores mencionados allí («*baina*»[2] y «*eta*»[3]) 5 marcadores del discurso ambiguos: «*berriz*»,[4] «*bait-*»,[5] «*-nez*»,[6] «*hala ere*»[7] y «*ere*».[8] Por lo tanto, hablamos de ambigüedad solamente si en el corpus analizado se ha empleado un marcador del discurso para señalar más de una relación retórica.

	DM	**Cant.**	**%**
Ambiguos	Eta	22	35,4
	Hala ere	4	6,4
	Bait-	3	4,8
	Baina	3	4,8
	Ere	3	4,8
	-nez	2	3,2
	Berriz	2	3,2
TOTAL AMBIGUOS	7 marcadores	39	62,9
TOTAL NO ANBIGUOS	20 marcadores	23	37,1

Tabla 5. Ambigüedad de los marcadores del discurso.

Este tipo de ambigüedad es bastante alta, ya que los 7 marcadores del discurso más empleados son los más ambiguos, en un 62,9 %. Y sólo podemos afirmar que en los otros 20 marcadores del discurso no hemos encontrado ambigüedad porque no hay más que una única instancia.

2 pero, mas, sino, aunque.
3 y/e, y en seguida, nada más, cuando, pues, porque, a causa de, con motivo de, pero, como, ya que.
4 en cambio, por el contrario, por su parte, a su vez.
5 porque, pues, ya que, que, el cual.
6 desde, puesto que, ya que.
7 a pesar de todo, sin embargo, además, por otra parte, por cierto.
8 también, tampoco, aun, incluso, hasta, ni, ni siquiera.

4. CONCLUSIONES

En este artículo nos hemos basado en la clasificación extendida de las relaciones retóricas (Mann y Taboada 2008) para analizar las relaciones de coherencia entre las unidades de 10 textos periódicos. Además hemos utilizado la herramienta RSTTOOL para el etiquetado del corpus.

El grado en el que se señalan las relaciones retóricas es bastante parecido en textos periodísticos en inglés (43,48 %) y en euskera (41,6 %). Existen 3 conductas posibles en la correspondencia entre las relaciones retóricas y los marcadores del discurso: concesión y contraste siempre señaladas; fondo, circunstancia, conjunción y elaboración señaladas algunas veces y la preparación sin señalar. Existe una pequeña diferencia en el orden canónico de las relaciones retóricas en inglés y en euskera. La relación de fondo tiene el orden inverso en euskera (núcleo-satélite). En un futuro será interesante analizar un corpus mayor y ver si existen más diferencias. De un modo general considerando las relaciones nucleares y multinucleares el 82,95 % de los marcadores del discurso han sido localizados en la segunda unidad. Se localizan más en las relaciones multinucleares (93,3 %) que en las unidades nucleares (74,2 %).

Además, hemos observado que la ambigüedad entre los marcadores del discurso es bastante alta: 62,9 %. No hemos apreciado correspondencia clara entre los marcadores del discurso y las relaciones retóricas, salvo en las relaciones de *concesión* y *contraste*. La relación de *concesión* es muy interesante ya que a aparecido marcada en todos los casos con algún marcador del discurso y además esos marcadores del discurso no han sido utilizados para realizar otras relaciones retóricas, lo que es equivalente a decir que tiene una función biyectiva entre las relaciones de *concesión* y los marcadores del discurso de concesión. En cambio, en la relación de *contraste*, no hay una relación biyectiva, aun estando marcada en todos los casos con marcadores del discurso de adversidad, los marcadores de adversidad también se utilizan para señalar otras relaciones; es por tanto una función sobreyectiva.

REFERENCIAS

Aduriz I. *et al.* 2006. «Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing». *Language and Computers* 56: 1-15.

Agirre E. *et al.* 2005. «EUSEMCOR: euskarako corpusa semantikoki etiketatzeko eskuliburua; editzatze-, etiketatze- eta epaitze-lanak». UPV/EHU/LSI/TR 23-2005.

Aldezabal I. *et al.* 2007a. «EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) segmentazio-mailan etiketatzeko eskuliburua». UPV/EHU/LSI/TR 11-2007.

Aldezabal I. *et al.* 2007b. «EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) dependentziekin etiketatzeko eskuliburua». UPV/EHU/LSI/TR 12-2007.

EUSKALTZAINDIA. 1990. *Euskal gramatika lehen urratsak – III (lokailuak)*. Bilbo.

EUSKALTZAINDIA. 2008. *Testu-antolatzaileak: erabilera estrategikoa*. Bilbo.

Iruskieta M., A. Diaz de Ilarraza y M. Lersundi. 2008. «Análisis de los marcadores del discurso para el euskera: denominación, clases, relaciones semánticas y tipos de ambigüedad». *XXVI Congreso Internacional de AESLA*. Almería.

Knott A. y R. Dale. 1994. «Using linguistic phenomena to motivate a set of coherence relations». *Discourse Processes* 18 (1): 35-62.

Mann W.C. y S.A. Thompson. 1987. «Rhetorical structure theory: A theory of text organization». Technical Report ISI/RS 87-190, ISI.

Mann W.C. y S.A. Thompson. 2000. «Toward a theory of reading between the lines: An exploration in discourse structure and implicit communication». *Proceedings of the 7th International Pragmatics Conference*. Budapest. Hungary.

Mann W.C. y M. Taboada. 2008. RST web-site: http://www.sfu.ca/rst/

Marcu M. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA: The MIT Press.

Martín Zorraquino M.A. y J. Portolés Lázaro. 1999. «Los marcadores del discurso». Eds. I. Bosque y V. Demonte. Vol. 3: 4051-4213.

O'Donnell, M. 1994. «RST-Tool: An RST Analysis Tool». *Proceedings of the 6th European Workshop on Natural language Generation*. Gerhard-Marcator University. Duisburg.

O'Donnell, M. 2000. «RSTTOOL 2.4 –A Markup Tool for Rhetorical Structure Theory». *Proceedings of the International Natural Language Generation Conference* (INLG'2000). Mitzpe Ramon. 253-256.

Portolés J. 1998. *Marcadores del Discurso*. Barcelona. Ariel.

Taboada M. 2006. «Discourse markers as signals (or not) of rhetorical relations». *Journal of Pragmatics* 38: 567-592.