Unidad discursiva y relaciones retóricas: un estudio acerca de las unidades de discurso en el etiquetado de un corpus en euskera

Discourse unit and rhetorical relations. A study about discourse units in the annotation of a corpus in Basque

Mikel Iruskieta
IXA NLP Group
Department of Didactics of Language and Literature
University of the Basque Country
Ramón y Cajal 72
48014 Bilbao
mikel.iruskieta@ehu.es

Arantza Díaz de Ilarraza
IXA NLP Group
Department of Computer Science
University of the Basque Country
Manuel Lardizabal 1
48014 Donostia
a.diazdeilarraza@ehu.es

Mikel Lersundi
IXA NLP Group
Department of Basque Philology
University of the Basque Country
Sarriena auzoa z/g
48940 Leioa
mikel.lersundi@ehu.es

Resumen: En este artículo se describe el estudio realizado sobre las características del etiquetado de la estructura de discurso, según la Teoría de la Estructura Retórica, en los niveles inter-oracional e intra-oracional. El corpus etiquetado está compuesto por textos médicos escritos en euskera y extraídos de la Gaceta Médica de Bilbao siendo nuestro objetivo final establecer una metodología general para la anotación de corpus a nivel discursivo. En este trabajo se analizan los acuerdos y desacuerdos de la anotación realizada por dos anotadores en cada nivel. Los resultados obtenidos sugieren que la segmentación en unidades de discurso es más compleja en el nivel intra-oracional mientras que la asignación de relaciones retóricas lo es en el nivel inter-oracional. Además hemos detectado que hay relaciones que aparecen con mayor frecuencia en cada nivel y otras se dan indistintamente en ambos niveles inter- e intra-oracional. Este estudio sienta las bases para el futuro desarrollo de un anotador automático de relaciones.
Palabras clave: anotación, análisis del discurso, segmentación, relaciones retóricas.

Abstract: This article describes the study on the features used for labelling the discourse structure, according to the Rhetorical Structure Theory, at the inter-sentential and intra-sentential levels. The tagged corpus is composed of medical texts written in Basque and extracted from the medical journal 'Gaceta Médica de Bilbao'. The difficulties encountered both while identifying the discourse units and while establishing the relations are analysed at each level based on the observation of agreement and disagreement identified in the texts annotated by two annotators. The results obtained suggest that the segmentation into units of discourse is more complex at the intra-sentential level while the assignment of rhetorical relations is more difficult at the inter-sentential level. We also note that some relations occur more frequently at the intra-sentential level and others at the inter-sentential level. However, there are relations that can appear indistinctively in both levels intra- and inter-sentential. This study will lay the foundations to carry out the automatic annotation process that the authors intend to perform shortly.
Keywords: Annotation, Discourse Analysis, Segmentation, Rhetorical Relations.

1 Introducción

El desarrollo de aplicaciones avanzadas basadas en el procesamiento del lenguaje, tales como búsqueda y extracción de información basada en conocimiento semántico, elaboración automática de resúmenes o traducción automática, precisan de corpus de referencia etiquetados a diferentes niveles lingüísticos: morfológico, sintáctico, semántico, etc. En este artículo trataremos del etiquetado de corpus a nivel discursivo.

La segmentación discursiva del corpus, al ser el primer estadio de la anotación de la estructura discursiva, tiene una importancia crucial y ha sido analizada desde diferentes puntos de vista y con finalidades diversas.

Existe una gran controversia sobre cuáles son los criterios de segmentación más adecuados para establecer las unidades de discurso. Cuando se trata de realizar la anotación discursiva de un corpus, normalmente se opta por realizar la segmentación considerando un alto nivel de granularidad estableciendo unidades de discurso a nivel intra-oracional (Carlson, Okurowski y Marcu 2002). El nivel inter-oracional (de menor granularidad) comprende unidades entre conjuntos de oraciones (párrafos, enunciados), unidades relacionadas mediante conjunciones coordinativas y unidades relacionadas de modo adverbial[1] (oraciones compuestas); en el nivel intra-oracional (de mayor granularidad) se consideran las unidades relacionadas con conjunciones subordinantes[2] y coordinativas (cláusulas con relaciones adverbiales); finalmente, el nivel de complementos verbales con sólo relaciones sintácticas hace referencia a los complementos de verbos declarativos, verbos que tienen como complementos otros verbos.

Cuando el objetivo es ofrecer un corpus de referencia enriquecido con información discursiva a la comunidad científica se suele optar por un alto nivel de granularidad, sin embargo Tofiloski, Brooke y Taboada (2009) subrayan que una granularidad tan fina, sobre los complementos de los verbos declarativos (*attributive and cognitive verbs*) no recoge información sobre relaciones retóricas, sino que recoge información de otras cuestiones del discurso. Limitándose a las relaciones retóricas Tofiloski, Brooke y Taboada (2009), descartan el último nivel (cláusulas con sólo relaciones sintácticas) y consideran únicamente niveles inter-oracional e intra-oracional. Esta distinción es útil, por ejemplo, para la clasificación de textos de diferentes géneros (Webber 2009); sin embargo, no lo es para tareas de resumen automático (Marcu 1999), donde es más conveniente considerar la granularidad inter-oracional.

En cuanto a la segmentación de alto nivel, Girill (1991) propone unidades discursivas más amplias (el pasaje) para tareas de recuperación de la información; en este sentido Hearst (1997) determina los multiparágrafos como unidades discursivas en la detección de cambios de tema.

Por tanto, del estudio bibliográfico se observa que la granularidad puede ser determinante para el éxito o no en ciertas tareas de etiquetado.

En este sentido, nuestro objetivo general es doble: i) establecer la metodología de anotación de la estructura relacional del discurso (anotación de segmentos y relaciones retóricas) y ii) llevar a cabo el proceso de anotación inter- e intra-oracional en un corpus.

De las diferentes teorías discursivas que formalizan la estructura referencial (Webber, et al 2003, Asher y Lascarides 2003, Polanyi 1988, Wolf y Gibson 2004), el marco teórico sobre el que desarrollamos este estudio empírico es la Teoría de la Estructura Retórica[3] (RST) de Mann y Thomson (1987), que es válida según Taboada y Mann (2006a) para aplicaciones avanzadas.

Con el fin de establecer la metodología de anotación nos preguntamos si existe el mismo grado de ambigüedad, en cuanto a las relaciones de la RST, en los niveles inter-oracional e intra-oracional. El objetivo concreto de este estudio es determinar, con la menor ambigüedad posible, el tipo de relaciones o las relaciones fácilmente identificables en cada nivel de manera que sirva como base en la implementación de un analizador automático de discurso.

Marcu y Echihabi (2002) sostienen que la anotación automática de ciertas relaciones retóricas conviene abordarla inicialmente en el nivel intra-oracional por ser el menos ambiguo. En la misma línea Soricut y Marcu (2003: 234) mencionan que algunas de las relaciones retóricas se derivan de las estructuras sintácticas:

> *Our experiments empirically show that, at the sentence level, there is an extremely strong correlation between syntax and discourse. This is even more remarkable given that the discourse*

[1] Thompson et al. (1985) detallan una tipología y sus funciones de oraciones adverbiales a ambos niveles: intra-oracional e inter-oracional.

[2] En este trabajo utilizamos el concepto de subordinación de modo tradicional. Véase Lehmann (1985) para una clasificación exhaustiva sobre los diferentes grados de dependencia entre sintagmas relacionales y un acercamiento funcional de la combinación de cláusulas.

[3] Página Web de la RST: http://www.sfu.ca/rst/

corpus (RST-DT, 2002) was built with no syntactic theory in mind. The annotators used by Carlson et al. (2003) were not instructed to build discourse trees that were consistent with the syntax of the sentences. Yet, they built discourse structures at sentence level that are not only consistent with the syntactic structures of sentences, but also derivable from them.

Pardo y Nunes (2008) han obtenido en la anotación de relaciones retóricas un grado más alto de acuerdo a nivel intra-oracional, en la evaluación de un analizador discursivo automático basado en patrones lingüísticos para el portugués de Brasil y en ese mismo nivel Soricut y Marcu (2003) han logrado para el inglés con un modelo estadístico un grado de robustez parecido al conseguido por anotadores humanos. Sin embargo, según Pardo y Nunes (2008), ese modelo estadístico de anotación no puede extenderse al nivel inter-oracional.

La estructura de este artículo es la siguiente: en la sección 2 explicamos el marco teórico y la metodología empleada para la anotación del corpus y su evaluación. Los resultados de las anotaciones de los niveles inter- e intra-oracional y su interpretación se presentan en las secciones 3 y 4 respectivamente. Finalmente, en la sección 5, establecemos las conclusiones y el trabajo futuro.

2 Teoría y metodología

2.1 Teoría

La RST es una teoría de carácter aplicado e independiente del idioma que nos permite describir la coherencia entre fragmentos textuales combinando la idea de nuclearidad, o importancia de un fragmento del discurso, con la identificación de las relaciones retóricas que unen los fragmentos del texto. Se entiende que el autor va guiando al lector, mediante el texto, comunicándole explícitamente o implícitamente qué fragmento es más importante y su relación con los demás fragmentos. Las relaciones se definen en base a las restricciones que se establecen entre el núcleo (N) y satélite (S), y el efecto que crea en el lector. Estas relaciones, según la teoría, pueden ser paratácticas (N-N), cuando se establece la relación entre fragmentos con el mismo grado de importancia en la intención del autor (LISTA, CONTRASTE, DISYUNCIÓN, etc.), o hipotácticas (N-S), cuando se establece una relación entre una unidad menos importante: satélite (S) con otra más importante: núcleo (N) siempre según la intención del autor. (ELABORACIÓN, MÉTODO, CONCESIÓN, CAUSA, RESULTADO, etc.). Las relaciones hipotácticas se clasifican en relaciones de presentación (P) y de contenido (C)[4].

Dado que éste es el primer estudio de estas características que se realiza para el euskara, nuestro objetivo es establecer las relaciones retóricas entre los fragmentos del discurso siguiendo las definiciones RST pero sin consensos previos ante las diferentes formas lingüísticas que señalan una u otra relación. Después estudiaremos las discrepancias y estableceremos los criterios lingüísticos que nos lleven a una anotación robusta. Por este motivo hemos elegido la clasificación extendida con 29 relaciones (Mann y Taboada 2010), dejando aparte las clasificaciones más complejas como por ejemplo la propuesta por Carlson, Marcu y Okurowski (2001) de 78 relaciones. Para la visualización y etiquetado de los fragmentos y relaciones hemos utilizado la herramienta RSTTOOL (O'Donnell 2000).

2.2 Metodología

La metodología de este estudio incluye tres fases.

1. Constitución del corpus. Se constituye el corpus que contiene todos los resúmenes en euskera extraídos de la Gaceta Médica de Bilbao[5] desde sus inicios hasta el año 2008. El corpus está compuesto por 20 documentos y tiene un tamaño de 3.024 palabras.

2. Niveles de anotación retórica. En primer lugar, tras un proceso en el que se establecieron unos criterios de anotación generales, dos anotadores segmentan los textos del corpus a nivel inter-oracional que comprende oraciones con verbo conjugado y después relacionan las unidades discursivas identificadas utilizando la clasificación RST extendida. En segundo lugar, se pide a los anotadores que vuelvan a segmentar de nuevo los textos del corpus, pero con una mayor granularidad, anotación intra-oracional, que comprende oraciones adverbiales en la misma oración. Para no repetir tareas, sólo se

[4] Véase su distribución en la Tabla 4.

[5] Los artículos se han extraído de la página Web de la revista Gaceta Médica de Bilbao: http://www.gacetamedicabilbao.org/web/es/.

relacionan los segmentos (*spans*[6]) intra-oracionales entre punto y punto.

3. Evaluación del etiquetado. Se evalúan y se comparan las anotaciones y se extraen conclusiones de las tareas de segmentación y del análisis retórico realizadas en ambos niveles: inter-oracional e intra-oracional. El método que hemos utilizado para evaluar las anotaciones retóricas es el propuesto por da Cunha e Iruskieta (2010). La Figura 1 y la Figura 2 ilustran los dos niveles de segmentación. La Figura 1 muestra un ejemplo de segmentación y su anotación retórica a nivel inter-oracional, donde se considera como unidad discursiva (EDU) aquélla que contiene un verbo conjugado, a excepción del título que constituye una EDU aunque no contenga verbo. En la Figura 2 se muestra un ejemplo de la segmentación y su anotación retórica sólo a nivel intra-oracional, donde se considera una unidad como EDU siempre que presente un verbo, conjugado o sin conjugar, sea o no subordinado.

En cuanto a la fase referente al establecimiento de las relaciones retóricas se ha pedido a cada anotador que primero relacione las unidades que van de punto a punto y después los párrafos de manera incremental y modular como propone Pardo (2005).

Observamos en los árboles que representan la anotación retórica a nivel inter-oracional, que hay un desacuerdo en la relación entre los *span*s 2 y 3; el anotador A1 detecta la presencia de una relación hipotáctica de ELABORACIÓN mientras que A2 anota una relación paratáctica de UNIÓN. Este desacuerdo en la interpretación está ligado al concepto de nuclearidad y es debido a la ausencia de elementos discursivos que faciliten la identificación de la relación retórica.

En la Figura 2 se representa la segmentación a nivel intra-oracional[7] del último segmento del ejemplo de la Figura 1. Se ha considerado la conjunción de cláusulas verbales con complementos, en la que hay un verbo no conjugado *aztertu* 'examinar' y otro conjugado *alderatu da* 'se ha comparado'. En este caso se establece la relación de SECUENCIA entre ambas cláusulas que se explicita mediante los verbos *aztertu* 'examinar' y *alderatu* 'comparar' y la conjunción *eta* 'y'.

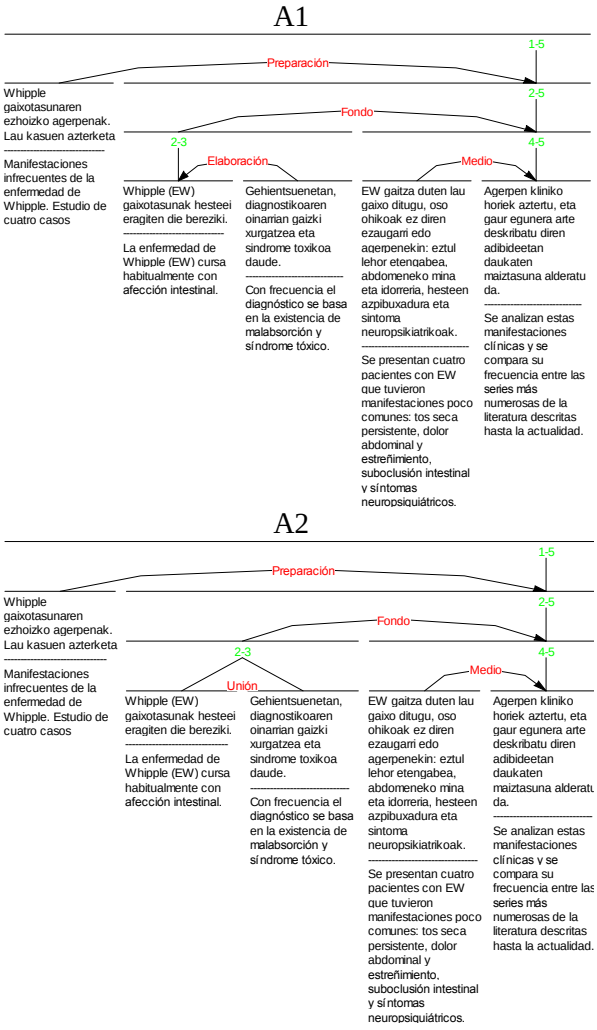

Figura 1: Anotación retórica inter-oracional

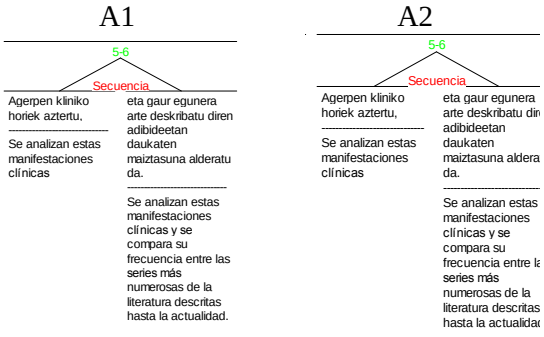

Figura 2: Anotación retórica intra-oracional

[6] El término *span* comprende tanto unidades de discurso como conjuntos de unidades.

[7] El fragmento de la Figura 2 es el correspondiente a la EDU 5 de la Figura 1. No hemos puesto las demás EDUs en la figura porque no reciben relación retórica alguna.

3 Resultados y discusión

En esta sección comparamos los resultados de las diferentes fases de cada tarea: segmentación y análisis retórico.

3.1 Segmentación

En cuanto a la evaluación de la segmentación se utilizan diferentes tipos de medidas: a) acuerdo promedio (*percent agreement*) (Marcu 1999, Hearst 1997, Passonneau y Litman 1993), que se utiliza para medir los posibles acuerdos entre anotadores; b) precisión y cobertura. Passoneau et al. (1993) lo utilizan para evaluar la fiabilidad del algoritmo de segmentación. Afantenos et al. (2010) utilizan *F-score*, medida utilizada en las tres tareas en CoNLL 2001, que combina ambas medidas (precisión y cobertura) para la anotación por pares de humanos y c) coeficiente Kappa. Esta otra medida que substrae el valor de la casualidad (Carletta 1996) es usada para medir el acuerdo entre anotadores en Hearst (1997) y Miltsakaki et al. (2004) y Tofiloski, Brooke y Taboada (2009), estos últimos comparan esta medida con *F-score*.

	EDU
Inter	100,00%
Intra	86,26%

Tabla 1: Acuerdo en la segmentación

En la Tabla 1 presentamos los resultados de la segmentación, donde podemos observar que el acuerdo nivel inter-oracional ha sido mayor que a nivel intra-oracional; ya que es mayor la complejidad de identificar las unidades a nivel intra-oracional por la variedad de casos que se presentan, sobre todo en una lengua aglutinante como el euskera.

3.2 Evaluación cualitativa del acuerdo en la anotación retórica

Con el método cuantitativo propuesto por Carlson et al. (2001) se mide el acuerdo en las anotaciones, dando especial importancia a la nuclearidad, donde se evalúa el acuerdo en los siguientes factores: i) segmentos simples del discurso (EDU), ii) segmentos compuestos, iii) nuclearidad y iv) relación. Aunque la nuclearidad es un factor de interés para muchas aplicaciones –por ejemplo, resumen automático (Marcu 1999) y detección del antecedente anafórico (Danlos 2008, Cristea, Ide y Romary 1998)–, no lo es para el objetivo concreto de este estudio. Nuestra propuesta es realizar una evaluación más cualitativa basada en los siguientes factores que intervienen en la asignación de la relación retórica: i) identificación de unidades núcleo a las que se asocia las relaciones o Asociación (A), ii) identificación de EDU o conjunto de *span*s de las unidades satélite (Composición: C) y iii) relaciones[8] (R). En la Tabla 2 se presenta el acuerdo en ambos niveles[9] atendiendo a los factores mencionados.

	A	**C**	**R**
Inter	71,23%	67,45%	57,00%
Intra	88,37%	92,06%	71,19%

Tabla 2. Cobertura en Asociación, Composición y Relación

Los resultados sugieren que la dificultad de la tarea a nivel inter-oracional es mayor, ya que a ese nivel hay menor acuerdo en todos los factores. Si tenemos en cuenta la cobertura es un 17,14% menor en la Asociación (A), un 24,61% en la Composición (C) y un 14,35% en la Relación (R). De estos datos deducimos que aunque la tarea de la segmentación es más compleja, el resto de las tareas a nivel intra-oracional son más simples. Las razones podrían ser las siguientes:

- La manera en que se combinan los segmentos es más simple a niveles más bajos (Composición).
- La identificación de las unidades núcleo a las que se asocian las relaciones es más sencilla a nivel intra-oracional que a nivel inter-oracional (Asociación).
- Tal y como apuntaban Marcu y Echihabi (2002) y Soricut y Marcu (2003), existe una fuerte relación entre sintaxis y discurso por lo tanto es más sencillo establecer la relación entre las unidades a nivel intra-relacional (Relación).

Nos fijamos ahora en los casos de acuerdo a nivel de relación y observamos más en detalle (Tabla 3) en qué casos se ha producido acuerdo total: i) acuerdo en Composición, Asociación y Relación (CAR); acuerdos parciales: ii) en Asociación y Relación (AR), iii) acuerdo en

[8] En da Cunha e Iruskieta (2010) se propone el modo de evaluar también la nuclearidad con este método cualitativo.

[9] La precisión es la misma para los tres factores. A nivel inter-oracional es de 100,00% y a nivel intra-oracional de 96,39%.

Composición y Relación (CR) y iv) acuerdo únicamente en Relación (R).

	CAR	AR	CR	R
Inter	83,60%	5,74%	4,10%	6,56%
Intra	93,10%	5,17%	1,72%	0,00%

Tabla 3. Tipos de acuerdo en base a la relación

Los resultados de la Tabla 3 sugieren que el acuerdo a nivel intra-oracional además de ser mayor es más consistente, ya que el acuerdo se basa en menor medida en acuerdos parciales (AR, CR y R).

3.3 Descripción de las relaciones retóricas

En este apartado se describe la frecuencia y ambigüedad de las relaciones hipotácticas en los niveles intra-oracional e inter-oracional.

R	Inter	Intra
Lista (N-N)	26,02%	15,52%
Elaboración (C)	21,95%	8,62%
Preparación (P)	17,07%	0,00%
Método (C)	10,57%	10,34%
Resultado (C)	8,94%	8,62%
Fondo (P)	8,13%	3,45%
Circunstancia (C)	0,00%	15,52%
Conjunción (N-N)	0,00%	8,62%
Condición (C)	0,00%	5,17%
Propósito (C)	0,00%	5,17%
Interpretación (C)	3,25%	3,45%
Concesión (P)	0,81%	3,45%
Evidencia (P)	0,81%	3,45%
Causa (C)	0,00%	3,45%
Contraste (N-N)	1,63%	1,72%
Justificación (P)	0,81%	0,00%
Motivación (P)	0,00%	1,72%
Secuencia (N-N)	0,00%	1,72%
Total	**100,00%**	**100,00%**

Tabla 4. Acuerdo en relaciones

En la Tabla 4 se presenta relación por relación[10] el acuerdo habido entre ambos anotadores en los diferentes niveles.

Considerando sólo los casos con un acuerdo superior al 5,00%, observamos que: i) a nivel intra-oracional, entre las relaciones hipotácticas, las más utilizadas y con mayor acuerdo, es decir, las menos ambigüas, son las relaciones de contenido (C) con formas subordinadas: CIRCUNSTANCIA, CONDICIÓN y PROPÓSITO; ii) a nivel inter-oracional, las relaciones de presentación (P): PREPARACIÓN y FONDO y iii) algunas relaciones de contenido (ELABORACIÓN, MÉTODO y RESULTADO) se utilizan en ambos niveles con frecuencia similar.

3.4 Descripción de la discrepancia

En relación con el desacuerdo de anotación las causas más discutidas son: a) la indeterminación de las relaciones retóricas por definición (Stede 2008), b) las diferentes y posibles interpretaciones (Taboada y Mann 2006b)[11] y c) falta de consenso previo (ver Tabla 5). Tras el estudio de las discrepancias se han detectado lo que podemos llamar patrones de confusión, que en nuestro caso se deben a los siguientes factores: a) segmentación, evidencia la dificultad de la segmentación a nivel intra-oracional; b) determinación de la nuclearidad; c) asignación de relaciones paratácticas, y d) asignación de relaciones hipotácticas.

Patrones de confusión	Inter	Intra
Segmentación	0,00%	27,00%
Nuclear (N-S)	54,00%	43,00%
Multinuclear (N-N)	13,00%	11,00%
Nuclear vs Multinuclear (N-S/N-N)	33,00%	19,00%

Tabla 5. Patrones de confusión de relaciones

Confusión en relaciones nucleares	Inter	Intra
Interpretación (C) / Resultado (C)	10,00%	0,00%
Justificación (P) / Causa (C)	1,00%	12,00%
Otras confusiones	43,00%	31,00%

Tabla 6. Patrones de confusión de relaciones

[10] En cada relación se especifica el tipo de relación. Según la RST hay dos tipos de relaciones hipotácticas: relaciones de presentación (P) y relaciones de contenido (C). Las otras relaciones son paratácticas o multinucleares (N-N).

[11] Aunque un texto puede tener más de una interpretación o árbol (Mann y Thompson 1987), se le ha pedido a los anotadores que den únicamente una interpretación.

Dentro de los patrones de confusión en relaciones hipotácticas es notable señalar que también los patrones de confusión (Tabla 6) son sensibles a nivel: INTERPRETACIÓN/RESULTADO a nivel inter-oracional y JUSTIFICACIÓN/CAUSA a nivel intra-oracional.

4 Conclusiones y trabajo futuro

Hemos presentado el estudio realizado sobre las características del etiquetado de la estructura de discurso, según la Teoría de la Estructura Retórica, en los niveles inter-oracional e intra-oracional. Este estudio nos sirve de base para establecer y refinar la metodología de etiquetado de estructuras de discurso. Basándonos en los resultados de este estudio indicamos que el acuerdo, en niveles más bajos del árbol retórico (nivel intra-oracional), es menor en la segmentación, un 13,74% menor; pero es mayor en la asignación de relaciones retoricas por su alto grado de señalización, un 14,35% mayor. Además, los resultados señalan que la configuración de las relaciones es diferente en un nivel u otro. Las relaciones hipotácticas en el nivel inter-oracional de mayor frecuencia y acuerdo son PREPARACIÓN y FONDO; mientras que en el nivel intra-oracional de las relaciones hipotácticas son: CIRCUNSTANCIA, CONDICIÓN y PROPÓSITO. Las relaciones con mayor acuerdo a nivel inter-oracional podrían ser explotadas en tareas de resumen automático y las del nivel intra-oracional en tareas de extracción de información. A pesar de los desacuerdos encontrados los resultados sugieren que la anotación automática de discurso debería considerar las tres relaciones intra-oracionales mencionadas por las siguientes razones: i) están siempre señalizadas y ii) ofrecen un bajo grado de ambigüedad. La identificación de estas relaciones nos puede servir de ayuda en el diseño de un anotador automático de relaciones retóricas. Además, en los patrones de confusión también identificamos claves importantes para dicho diseño en lo referente a las relaciones retóricas INTERPRETACIÓN/RESULTADO a nivel inter-oracional y JUSTIFICACIÓN/CAUSA a nivel intra-oracional.

En trabajos futuros analizaremos las razones lingüísticas de la correlación entre sintaxis y discurso en la anotación automática de relaciones retóricas y abordaremos las razones de los patrones de confusión para realizar árboles de decisiones o manual detallado de las marcas que evidencian las relaciones.

Agradecimientos

Este trabajo ha sido realizado en el marco de los siguientes proyectos: Grupo IXA, Grupo consolidado 2007-2012 (IT-397-07) [Gobierno Vasco]; KNOW2 (TIN2009-14715-C04-01) [MICCIN], Hibrido Sint (TIN2010-20218) [MICCIN], y GARATERM2 (US10/01) [Gobierno Vasco].

Bibliografía

Afantenos, S., P. Denis, P. Muller y L. Danlos, 2010. Learning Recursive Segments for Discourse Parsing. En *Proceedings of the Seventh conference on International Language Resources and Evaluation*, 3578-3584.

Asher, N. y A. Lascarides, 2003. *Logics of conversation.* Cambridge Univ Pr, Cambridge.

Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22 (2): 249-254.

Carlson, L., D. Marcu y M.E. Okurowski, 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. En *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 85-112.

Carlson, Lynn, M.E. Okurowski, D. Marcu, 2002. RST Discourse Treebank. *LDC*.

Cristea, D., N. Ide y L. Romary, 1998. Veins theory: A model of global discourse cohesion and coherence. En *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, 281-285.

da Cunha, I. y M. Iruskieta, 2010. Comparing rhetorical structures in different languages: The influence of translation strategies. *Discourse Studies*, 12 (5): 563-598.

Danlos, L., 2008. Strong generative capacity of RST, SDRT and discourse dependency DAGSs. *Constraints in discourse*, 69–95.

Girill, T., 1991. Information chunking as an interface design issue for full-text databases. *Interfaces for Information Retrieval and Online Systems: The State of the Art*, 149-158.

Hearst, M.A., 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23 (1): 33-64.

Lehmann, C., 1985. Towards a typology of clause linkage. En *Conference on Clause Combining*, 181-248.

Mann, W.C. y M. Taboada, 2010. RST web-site. http://www.sfu.ca/rst/.

Mann, W.C.y S.A. Thompson, 1987. Rhetorical Structure Theory: A Theory of Text Organization. Marina del Rey. CA: Information Sciences Institute.

Marcu, D., 1999. Discourse trees are good indicators of importance in text. *Advances in automatic text summarization*, 123-136.

Marcu, D. y A. Echihabi, 2002. An unsupervised approach to recognizing discourse relations. En *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 368-375.

Miltsakaki, E., R. Prasad, A. Joshi y B. Webber, 2004. Annotating discourse connectives and their arguments. En *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*, 9–16.

O'Donnell, M., 2000. RSTTool 2.4: a markup tool for Rhetorical Structure Theory. En *Proceedings of the First International Conference on Natural Language Generation INLG '00*, 253-256.

Pardo, T.A.S. y M.G.V. Nunes, 2008. On the development and evaluation of a Brazilian Portuguese discourse parser. *Journal of Theoretical and Applied Computing*, 15 (2): 43-64.

Passonneau, R.J. y D.J. Litman, 1993. Intention-based segmentation: Human reliability and correlation with linguistic cues. En *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, 148-155.

Polanyi, L., 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, 12 (5-6): 601-638.

Soricut, R. y D. Marcu, 2003. Sentence level discourse parsing using syntactic and lexical information. En *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 149-156.

Stede, M., 2008. Disambiguating rhetorical structure. *Research on Language & Computation*, 6 (3): 311-332.

Taboada, M. y W.C. Mann, 2006a. Applications of rhetorical structure theory. *Discourse studies*, 8 (4): 567.

Taboada, M. y W.C. Mann, 2006b. Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies*, 8 (3): 423.

Thompson, S.A., R. Longacre y S.J.J. Hwang, 1985. Adverbial clauses. En: Shopen,T. (Ed.), Language Typology and Syntactic Description: Complex Constructions. Cambridge University Press, New York : 171-234.

Tofiloski, M., J. Brooke y M. Taboada, 2009. A syntactic and lexical-based discourse segmenter. En *Proceedings of the ACL-IJCNLP 2009*, 77-80.

Webber, B., 2009. Genre distinctions for discourse in the Penn TreeBank. En *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, 674-682.

Webber, B., M. Stone, A. Joshi y A. Knott, 2003. Anaphora and discourse structure. *Computational Linguistics*, 29 (4): 545-587.

Wolf, F. y E. Gibson, 2004. Representing discourse coherence: A corpus-based analysis. En *Proceedings of the 20th international conference on Computational Linguistics*, 134-140.