Metodología y desarrollo del primer corpus en español anotado con relaciones retóricas

Methodology and development of the first corpus in Spanish annotated with rhetorical relations

Iria da Cunha
Institut Universitari de Lingüística Aplicada (UPF, España), Universidad Nacional Autónoma de México (IINGEN, México), Laboratoire Informatique d'Avignon (UAPV, Francia)
C/Roc Boronat, 138, 08018, Barcelona
iria.dacunha@upf.edu

Juan-Manuel Torres-Moreno
Laboratoire Informatique d'Avignon (UAPV, Francia), Universidad Nacional Autónoma de México (IINGEN, México), École Polytechnique de Montréal (Canadá)
339, chemin des Meinajaries, Agroparc, 84911 Avignon
juan-manuel.torres@univ-avignon.fr

Gerardo Sierra
Universidad Nacional Autónoma de México (IINGEN, México)
Circuito Escolar s/n, Ciudad Universitaria, Coyoacán,
México D.F. 04510
gsierram@iingen.unam.mx

Resumen: En este trabajo mostramos la metodología empleada para el desarrollo del primer corpus de textos en español anotados con las relaciones retóricas de la *Rhetorical Structure Theory*. Detallamos cuestiones relacionadas con la selección de los textos, la teoría empleada para anotarlos, el diseño de la interfaz, la selección y entrenamiento de los anotadores, el diseño y la gestión del procedimiento de anotación, la validación de los resultados, y la difusión y el mantenimiento del producto.
Palabras clave: corpus, textos especializados, relaciones retóricas, discurso, anotación

Abstract: In this work we show the methodology used in order to carry out the development of the first corpus including texts in Spanish annotated with the rhetorical relations of the Rhetorical Structure Theory. We detail some issues related with texts' selection, theory used for the annotation, interface designing, selection and training of annotators, designing and managing the annotation procedure, validation of results, and delivering and maintaining of the product.
Keywords: corpus, specialized texts, rhetorical relations, discourse, annotation

1 Introducción

La *Rhetorical Structure Theory* (RST) de Mann y Thompson (1988) es una teoría independiente de la lengua que parte de la idea de que un texto puede segmentarse en Unidades Discursivas Mínimas (EDUs) vinculadas mediante relaciones retóricas[1] núcleo-satélite o multinucleares. En las primeras, el satélite aporta una información adicional sobre el núcleo (ej. Resultado, Condición o Concesión); en las segundas, diversos núcleos están conectados al mismo nivel, es decir, no hay elementos dependientes de otros (ej. Contraste, Lista o Secuencia). Así, el árbol discursivo de la Figura 1 contiene una relación núcleo-satélite (de Justificación) y 2 relaciones multinucleares (de Conjunción y de Contraste).

[1] En este trabajo empleamos los términos "discursivo" y "retórico" como sinónimos.

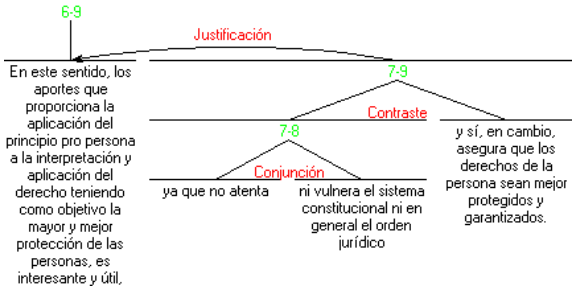

Figura 1: Ejemplo de árbol retórico de la RST

Esta teoría ha resultado de gran utilidad para desarrollar aplicaciones de diversos tipos, como resumen automático, extracción de información, generación de texto, traducción automática, etc. (Taboada y Mann, 2006). Sin embargo, la mayor parte de estas investigaciones se realizaron para el inglés, el alemán o el portugués, ya que en estas lenguas existen corpus anotados con relaciones de la RST (véase apartado 2) y 2 de ellas disponen de analizadores automáticos que emplean esta teoría (en inglés: Marcu, 2000; en portugués: Pardo et al., 2008). Sin embargo, la comunidad científica que trabaja sobre la RST aplicada al español es muy reducida (ej. Bouayad-Agha et al., 2006; da Cunha et al., 2007; da Cunha e Iruskieta, 2010; Romera, 2004; Taboada, 2004). En este contexto, consideramos necesaria la construcción de un corpus en español anotado con las relaciones retóricas de la RST, que permita el desarrollo de un analizador discursivo automático para esta lengua y de diversas aplicaciones relacionadas con la lingüística computacional como las desarrolladas en otras lenguas (traducción automática, resumen automático, extracción de información, etc.). Precisamente, este es el objetivo de nuestro trabajo. En este artículo mostramos el estado del primer corpus de este tipo para el español, denominado RST Spanish Treebank, el cual estamos desarrollando en la actualidad.

En el apartado 2 mostramos trabajos previos sobre corpus anotados con la RST. En el apartado 3 explicamos la metodología empleada para el desarrollo del corpus. En el apartado 4 establecemos algunas conclusiones y en el apartado 5 detallamos el trabajo futuro.

2 Antecendentes

El principal corpus anotado mediante la RST es el RST Discourse Treebank, para el inglés (Carlson et al., 2002). Incluye 385 textos del dominio periodístico, extraídos del Penn Treebank (Marcus et al., 1993), como reseñas culturales, cartas al director, editoriales, reportajes, etc. Está dividido en 347 textos como corpus de aprendizaje y 38 como corpus de prueba. Contiene 176.389 palabras y 21.789 EDUs. Un 13,8% de los textos fueron etiquetados por 2 personas. Se emplearon 78 relaciones. La herramienta empleada para la anotación fue la RSTtool (O'Donnell, 2000). Las principales ventajas de este corpus son que su tamaño es elevado y que la metodología de anotación es clara (detallada en el manual de anotación de Carlson y Marcu, 2001). La desventaja es que no es gratuito, no está en línea, no dispone de una interfaz de consulta y solo incluye textos del dominio periodístico.

El Discourse Relations Reference Corpus (Taboada y Renkema, 2008), también en inglés, incluye 65 textos de diversos tipos y fuentes: 21 artículos del Wall Street Journal extraídos del RST Discourse Treebank, 30 críticas de películas y libros extraídos del sitio web epinions.com, y 14 textos variados, incluyendo cartas, webs, artículos de revista, editoriales de periódico, etc. La herramienta empleada para la anotación también fue la RSTtool. Sus ventajas son que es gratuito, se puede descargar en línea e incluye textos muy variados. Las desventajas son que el número de textos no es muy elevado, no se explicita la metodología de anotación y no incluye textos anotados por diversas personas.

El Potsdam Commentary Corpus, elaborado para el alemán (Stede, 2004; Reitter y Stede, 2003), incluye 173 textos sobre política del periódico en línea Märkische Allgemeine Zeitung, 32.962 palabras y 2.195 oraciones. Se anotó con diversas informaciones: morfología, sintaxis, estructura retórica, conectores, correferencia y estructura informativa. Sin embargo, solo una parte de este corpus (10 textos), lo que los autores denominan "core corpus", está anotado con todas estas informaciones. Los textos se anotaron con la RSTtool. Las ventajas de este corpus son varias: está anotado a diversos niveles (es especialmente interesante la anotación automática de conectores, que evidencian relaciones retóricas); todos los textos fueron anotados por 2 personas (que pasaron una fase previa de entrenamiento sobre la RST); se puede solicitar a los creadores el envío del

corpus de manera gratuita si su uso es con fines de investigación, y existe una herramienta de consulta del corpus que permite realizar búsquedas entre niveles de anotación. Las desventajas son que los textos del corpus solo se corresponden con un género discursivo y un ámbito temático, la metodología de anotación fue bastante intuitiva (sin un manual), no se ofrece el porcentaje de acuerdo entre anotadores y la herramienta de consulta del corpus no está disponible en línea.

Para el portugués, existen 2 corpus, realizados con el fin de desarrollar un analizador discursivo automático. El primero, el CorpusTCC (Pardo et al., 2008), sirvió como corpus de aprendizaje para la detección de patrones lingüísticos que evidenciasen relaciones retóricas. Contiene 100 secciones de introducción de tesis de informática, que suponen 53.000 palabras y 1.350 oraciones. Se usaron 32 relaciones retóricas. Se siguió el manual de anotación de Carlson y Marcu (2001), adaptado al portugués. La herramienta de anotación fue la ISI RST Annotation Tool[2], una extensión de la RSTtool. Las ventajas de este corpus es que contiene un número aceptable de textos y palabras, sigue una metodología de anotación específica y se puede descargar gratuitamente. Las desventajas son que incluye textos de un solo género y dominio, está anotado solo por una persona y no dispone de interfaz de consulta.

El segundo corpus en portugués, Rhetalho (Pardo y Seno, 2005), se empleó como corpus de referencia para la evaluación del analizador. Contiene 50 textos: 20 secciones de introducción y 10 de conclusiones de artículos científicos de informática, y 20 textos del periódico en línea Folha de São Paulo (7 de la sección de Cotidiano, 7 de Mundo y 6 de Ciencia). Incluye aproximadamente 5.000 palabras. Las relaciones anotadas y la herramienta de anotación son las mismas que las empleadas en el CorpusTCC. Las ventajas de este corpus son que fue anotado por 2 personas expertas en la RST a las que se ofreció un protocolo de anotación para posteriormente obtener el acuerdo entre ellos, contiene textos de géneros y dominios diferentes, y se puede descargar gratuitamente. Las desventajas son que el tamaño es muy reducido y no dispone de interfaz de consulta.

[2] http://www.isi.edu/~marcu/discourse/

En la Tabla 1 se resumen las características de estos 5 corpus.

	RST Dis. Tree.	**Dis. Ref. Cor.**	**Pots. Cor.**	**TCC**	**Rhetallo**
Lengua	inglés	inglés	alemán	portug.	portug.
Tamaño	385 text. 176.389 pal.	65 text.	173 text. 32.962 pal.	100 text. 53.000 pal.	50 text. 5.000 pal.
Tema	diversos	diversos	política	infor-mática	diversos
Género	diversos	diversos	artículo perio-dístico	intro-ducción de tesis	artículo científico, artículo periodís-tico
Doble anotación	sí (53 textos)	no	sí (todos los textos)	no	sí (todos los textos)
Protocolo anotación	sí	no	no	sí	sí
Herra-mienta anotación	RSTtool	RSTtool	RSTtool	ISI RST Annot. Tool	ISI RST Annot. Tool
Interfaz consulta	no	no	sí (no en línea)	no	no
Acceso	de pago	gratuito (descar-ga en línea)	gratuito (bajo peti-ción)	gratuito (descar-ga en línea)	gratuito (descarga en línea)

Tabla 1: Resumen de las características de los corpus existentes anotados con la RST

3 Metodología y desarrollo del corpus

Como Sierra (2008: 446) afirma, "un corpus lingüístico consiste en la recopilación de un conjunto de textos de materiales escritos y/o hablados, agrupados a partir de un conjunto de criterios mínimos, para realizar ciertos análisis lingüísticos". Hovy (2010) menciona 7 cuestiones relevantes a la hora de elaborar un corpus, a las que nos referiremos en los siguientes subapartados.

3.1 Selección del corpus

Para que un corpus pueda considerarse representativo debe incluir textos de diversos tipos. Con este objetivo, el RST Spanish Treebank incluye textos de diversos géneros, temáticas, fuentes, autores, etc. (resúmenes de artículos en actas de congresos y en revistas científicas o de divulgación; apartados de tesis

doctorales, de artículos de investigación, de páginas webs de asociaciones o instituciones; reportajes de revista; fragmentos de libros de texto escolares). Todos los textos del corpus son especializados, considerando como tales los textos escritos por profesionales del dominio del que trata un texto (Cabré, 1999). Los textos especializados pueden dividirse en 3 niveles: alto (emisor y receptor especialistas), medio (emisor especialista y receptor aprendiz) y bajo (emisor especialista y receptor lego). Así, el corpus incluye textos especializados de los 3 niveles: alto (artículos científicos, actas de congresos, tesis doctorales, etc.), medio (libros de texto escolares) y bajo (artículos y reportajes de divulgación, webs de asociaciones, etc.). Los textos se han dividido en 9 ámbitos temáticos (algunos de los cuales contienen subdivisiones). En la Tabla 2 se incluye el número de textos y palabras por ámbito. Algunos ámbitos incluyen un número mayor de textos que otros. Esto no es sorprendente ya que "en el caso de un corpus especializado, por ejemplo, es esperable que contenga más textos de un área que de otra" (Sierra, 2008: 450-451). El corpus total contiene 52.746 palabras (el texto más largo contiene 1.051 palabras y el más corto 25) y 267 textos. En la Tabla 3 se incluyen las estadísticas en cuanto a textos, palabras, oraciones y EDUs.

Ámbito	**Nº de textos**	**Nº de palabras**
1. Astrofísica	**17**	**2.064**
2. Derecho	**3**	**532**
3. Economía	**16**	**2.531**
4. Ingeniería sísmica	**3**	**676**
5. Lingüística	**46**	**11.817**
Adquisición del lenguaje	17	1.931
Lingüística aplicada	11	3.122
Procesamiento del Lenguaje Natural	3	441
Terminología	15	6.323
6. Matemáticas	**78**	**11.256**
Primaria	15	3.489
Bachillerato	15	3.503
Artículos científicos	48	4.264
7. Psicología	**31**	**4.963**
8. Sexualidad	**41**	**11.698**
Perspectiva clínica	26	9.264
Nivel alto	10	3.788
Nivel bajo	16	5.476
Perspectiva psicológica	15	2.434
9. Medicina	**32**	**7.209**
Oncología	9	2.380
Administración de servicios de salud	20	4.002
Ortopedia	3	827

Tabla 2: Número de textos y palabras por ámbito

	Textos	**Palabras**	**Oraciones**	**EDUs**
Corpus de aprendizaje	183	41.555	1.759	2.655
Corpus de prueba	84	11.191	497	694
Corpus TOTAL	267	52.746	2.256	3.349

Tabla 3: Estadísticas del corpus

Sin embargo, como afirma Sierra (2008: 452), "resulta absurdo tratar de construir un corpus exhaustivo que cubra todos los aspectos de la lengua. Por el contrario, el lingüista busca la representatividad en los textos, esto es, intenta recoger una muestra de la lengua que se estudia para seleccionar los ejemplos cercanos a la realidad lingüística con el fin de analizarlos de manera pertinente". En este sentido y en el marco de este trabajo, más que medir el corpus en número de palabras o de textos, consideramos que su tamaño será adecuado si incluye un número representativo de ejemplos de relaciones discursivas, al menos 20 casos de cada una de ellas. En el Anexo 1 se muestra el número de ocurrencias de relaciones de cada tipo existentes actualmente en el corpus (N-S: relación núcleo-satélite; N-N: relación multinuclear). Como puede observarse, cuenta con más de 20 ejemplos para cada relación, excepto en las relaciones núcleo-satélite de Capacitación, Evaluación, Resumen, Alternativa y Unless, y las relaciones multinucleares de Disyunción y Conjunción, ya que no es tan habitual encontrar este tipo de relaciones retóricas en la lengua, en comparación con otras.

3.2 Marco teórico

Los criterios empleados para la segmentación y anotación son similares a los originales de Mann y Thompson (1988) para el inglés, y a los empleados en el trabajo para el español de da Cunha e Iruskieta (2010). Exploramos también el manual de anotación del inglés de Carlson y Marcu (2001), y, aunque empleamos algunos de sus postulados (como la no-relación Same-Unit), consideramos que el análisis que ellos realizan es demasiado minucioso en algunos aspectos y no se adaptan a nuestros intereses: encontrar un método de anotación lo más sencillo y objetivo posible, orientado al desarrollo de un analizador discursivo para el español basado en esos criterios. Resumidamente, nuestros criterios de segmentación, que se aplican manualmente, son los siguientes:

a) En primer lugar, se realiza una segmentación de todas las oraciones del texto (se entiende como oración un fragmento textual que comprende de punto a punto, punto y coma, interrogante o exclamación, contenga o no verbo; también se segmentan los títulos de los textos).

b) En segundo lugar, se observa si es posible segmentar EDUs a nivel intraoracional, dependiendo de los siguientes criterios:

b1) Toda EDU intraoracional debe contener un verbo conjugado, en infinitivo o en gerundio.

b2) No se consideran como EDUs las clásulas subordinadas de sujeto, de complemento directo, de complemento indirecto o completivas.

b3) No se consideran como EDUs las clásulas subordinadas de relativo.

b4) Los elementos entre paréntesis solo se segmentan si siguen el criterio b1).

b5) Las unidades imbricadas se segmentan mediante la no-relación Same-Unit.[3]

[3] La no-relación Same-Unit fue propuesta por Carlson y Marcu (2001), para dar cuenta de una EDU que aparece cortada por incluir otra EDU. En el siguiente ejemplo, las unidades 1 y 3 conforman una única EDU, que a su vez contiene otra EDU, formada por la unidad 2: [Se ha puesto de manifiesto,]1 [y así se ha atestiguado,]2 [una tendencia a importar préstamos.]3

3.3 Interfaz

La herramienta de anotación empleada en este trabajo es la RSTtool, una interfaz gratuita, amigable y sencilla. Así, no hemos diseñado una herramienta de anotación, sino que hemos empleado una ya existente. Sin embargo, sí hemos desarrollado una interfaz en línea[4] para albergar el corpus. Esta interfaz permite la visualización y descarga de los textos del corpus en txt, con su correspondiente árbol anotado en el formato de la RSTtool (rs3) y en formato imagen (png). De cada texto se ofrece su título, su referencia bibliográfica, su enlace web si es un documento en línea y su número de palabras. La interfaz muestra la división de los textos por ámbitos y permite seleccionar un subcorpus de interés para el usuario (formado por archivos individuales y/o carpetas que incluyen diversos archivos), que puede guardarse en local (generando un archivo xml) para futuros análisis. La interfaz cuenta con 4 herramientas de análisis automático desarrolladas en perl. Tres de ellas ofrecen estadísticas sobre el subcorpus seleccionado, con respecto a cantidad de: a) palabras, b) EDUs y c) relaciones discursivas. La RSTtool ofrece la funcionalidad c), pero solo permite aplicarla sobre un texto. Consideramos que obtener estadísticas sobre un corpus de textos es más útil para obtener resultados estadísticos significativos. Como la RSTtool, la herramienta permite contabilizar las relaciones multinucleares de 2 formas: a) 1 unidad por cada relación multinuclear o b) 1 unidad por cada núcleo. Así, el siguiente fragmento, que incluye 3 núcleos de Unión, puede contabilizarse como 1 ó 3 unidades, dependiendo de si se emplea a) o b), respectivamente:

[Su material genético es RNA diploide de sentido positivo,]$_{N\_Unión}$ [posee envoltura]$_{N\_ Unión}$ [y tiene un diámetro de 80 a 120 nm.]$_{N\_ Unión}$

El número de relaciones mostrado en el Anexo 1 se ha contabilizado mediante la estrategia a). Mediante la estrategia b), este número cambia y se obtienen 864 relaciones de Lista, 537 de Unión, 289 de Secuencia, 153 de Contraste, 28 de Conjunción y 24 de Disyunción.

La cuarta herramienta permite al usuario seleccionar un subcorpus y extraer de él los

[4] http://corpus.iingen.<unam.mx/rst/

satélites (fragmentos textuales) que se corresponden con la relación retórica que desee. Los resultados se ofrecen en un único documento, con lo cual esta herramienta podría considerarse también como un resumidor multidocumento guiado por el interés del usuario.

La interfaz cuenta con una pantalla para el envío de textos anotados por parte de los usuarios. Nuestra intención es que este corpus sea dinámico y sean los propios usuarios quienes puedan alimentarlo. Esta visión tiene una doble ventaja ya que, por un lado, el corpus crecerá y, por otro lado, los usuarios podrán beneficiarse de las aplicaciones de la interfaz, usando sus propios subcorpus. El único requisito que se solicita es que los textos estén anotados con las relaciones y criterios de segmentación y anotación de nuestro proyecto. Una vez enviados los textos, desde la coordinación se revisará la adecuación de las anotaciones a estos criterios.

3.4 Anotadores

Con respecto a los anotadores del corpus, se contó con un equipo de 10 personas (estudiantes de último curso de licenciatura, estudiantes de máster y doctores), que previamente a la anotación siguieron un curso de 6 meses (100 horas) sobre la RST y la metodología de segmentación y anotación empleada para el desarrollo del RST Spanish Treebank. Hemos denominado este período como “fase de entrenamiento”. El curso constó de un parte de teoría y de una parte práctica. En la parte teórica se ofreció a los anotadores una serie de criterios con respecto a las 3 fases del análisis discursivo: segmentación, detección de relaciones y construcción de árboles discursivos. En la parte práctica, en primer lugar se familiarizó a los anotadores con la RSTtool. En segundo lugar, se segmentaron textos extraídos de múltiples fuentes por los anotadores (siguiendo sus intereses personales, por ejemplo: webs de música, de videojuegos, de cocina, de arte, etc.), siguiendo los criterios de segmentación establecidos. Una vez segmentados, se pusieron en común dudas y problemáticas, y se intentó llegar a un acuerdo sobre los casos más complicados. En tercer lugar, en base a una lista cerrada de relaciones, se analizaron las relaciones existentes en los textos y, una vez más, se pusieron dudas en común para solucionar casos polémicos y reanotar los textos. Este proceso fue doblemente interesante, ya que, por un lado, permitió a los anotadores formarse y tener unos criterios comunes para la anotación del corpus, y, por otro lado, fue útil para definir los criterios de anotación de una manera más clara y consensuada, de cara a la anotación de los textos incluidos finalmente en el corpus.

3.5 Procedimiento de anotación

Una vez pasada la fase de entrenamiento de los anotadores y seleccionados los textos incluidos en el corpus, se pasó a la “fase de anotación”. En esta fase se asignaron los textos a los anotadores y se les indicó que debían realizar la anotación de manera individual y sin consultas entre ellos. Una vez segmentado un texto, se anotan las relaciones retóricas entre las EDUs. Primero se relacionan las EDUs dentro de una misma oración de manera binaria; después, las oraciones que forman parte del mismo párrafo; finalmente los párrafos entre ellos. La lista de relaciones retóricas empleadas se muestra en el Anexo 1. Siguiendo la metodología empleada en el RST Discourse Treebank, empleamos una parte de los textos como corpus de aprendizaje y otra como corpus de prueba. Concretamente, usamos un 69% de los textos como corpus de aprendizaje (183 textos) y un 31% como corpus de prueba (84 textos). Los textos del corpus de aprendizaje fueron anotados por 1 persona, mientras que los textos del corpus de prueba fueron anotados por 2 personas.

3.6 Validación de resultados

El acuerdo entre los anotadores de los textos del corpus de prueba se midió con las medidas de precisión y cobertura (una anotación respecto a la otra) mediante la herramienta automática en línea de comparación de árboles de la RST, RSTeval, de Mazeiro y Pardo (2009). En esta herramienta se ha implementado para 4 lenguas (inglés, portugués, español y euskera) la metodología de comparación de árboles de Marcu (2000), que evalúa la coincidencia de 4 aspectos: las unidades discursivas mínimas (EDUs), los conjuntos de EDUs (SPANs), la aparición de núcleos o satélites (Nuclearidad) y las relaciones discursivas anotadas (Relaciones). Esta metodología ha sido expresamente diseñada para comparar de

manera adecuada la coincidencia de estos elementos discursivos incluidos en dos anotaciones realizadas sobre el mismo texto; además, ha sido empleada previamente por otros autores (da Cunha e Iruskieta, 2010) para comparar árboles retóricos. Es por este motivo que no hemos empleado otras medidas de acuerdo entre anotadores, como kappa.

Tomando como referencia las anotaciones de uno de dos anotadores (elegido al azar), se midió la precisión y cobertura de cada par de árboles del corpus de prueba por separado; después se sumaron los resultados para obtener cifras generales. En la Tabla 4 se muestran los resultados globales para cada categoría. La categoría en la que más acuerdo hubo fue la de EDUs, que se corresponde con la segmentación. Esta cifra era esperable, ya que los criterios de segmentación ofrecidos a los anotadores eran precisos y la posibilidad de error era reducida. El acuerdo más bajo se obtuvo en la categoría de Relaciones. Consideramos que esta cifra, aunque es menor que la obtenida en la segmentación, es respetable y puede considerarse que hay un acuerdo elevado. En el RST Discourse Treebank se midió el acuerdo entre anotadores en diversas fases del proyecto. En la primera fase (antes de refinar su manual de anotación) obtuvieron los valores de kappa siguientes: 0.87 en EDUs, 0.77 en spans, 0.70 en nuclearidad y 0.60 en relaciones. Se observa entonces que la tendencia fue similar a la detectada en nuestro corpus: el mayor acuerdo se obtiene al nivel de la segmentación y el menor en la detección de relaciones.

Cateogoría	Precisión	Cobertura
EDUs	87,20%	91,04%
SPANs	86%	87,31%
Nuclearidad	82,46%	84,66%
Relaciones	76,81%	78,48%

Tabla 4: Acuerdo entre anotadores

Una vez observadas las cifras del acuerdo, analizamos las principales causas de discrepancia entre anotadores, que fueron: con respecto a la segmentación, error humano y falta de especificación de algunas cuestiones en el protocolo; con respecto a la anotación, ambigüedad de algunas relaciones (como Justificación y Causa, o Antítesis y Concesión) y diferencias al determinar la nuclearidad.

3.7 Difusión y mantenimiento

Los derechos de autor de los textos que se incluyen en un corpus es un tema polémico. Normalmente debe solicitarse autorización escrita a los autores de los textos para poder incluirlos en un corpus. Sin embargo, “existen excepciones o límites a las normas para las que no se requiere solicitar autorización del uso de la obra. Una de ellas es cuando se trata de trabajos de investigación o docencia y sin fines de lucro, siempre y cuando no se permita el acceso más allá de fragmentos de textos y se indique claramente la procedencia de los mismos” (Sierra, 2008: 452). Este es precisamente el caso de este corpus, ya que se trata de un proyecto de investigación con el que no pretendemos lucrarnos, solo ofrecemos fragmentos de textos más amplios (por ejemplo, el resumen de un artículo científico, un apartado de una página web, una introducción de una tesis, etc.) y todos los textos incluyen su correspondiente referencia bibliográfica, además del enlace al sitio web del texto si se trata de una publicación electrónica.

La descripción del corpus anotado es también una cuestión muy importante (Ide y Pustejovsky, 2010). Es necesario aportar una descripción detallada del corpus, que incluya el marco teórico, la metodología, los medios de mantenimiento del recurso, los aspectos técnicos, el director del proyecto, el equipo, etc. Nuestro corpus incluye toda esta información de manera detallada.

4 Conclusiones

Consideramos que este trabajo supone un paso importante para el estudio de la RST en español y que este corpus será de utilidad para realizar diversas investigaciones sobre el tema en esta lengua, tanto desde un punto de vista descriptivo (análisis contrastivos de textos de dominios especializados diferentes, análisis de géneros discursivos, análisis de marcadores discursivos, etc.) como aplicado (desarrollo de analizadores discursivos automáticos, desarrollo de aplicaciones de lingüística computacional, como resumen automático, traducción automática, extracción de información, etc.). Actualmente el tamaño del corpus es aceptable y, aunque el porcentaje de textos doblemente anotados es reducido, consideramos que el hecho de contar con 10

anotadores (que siguen un mismo protocolo de anotación) evita que la anotación del corpus esté sesgada. Además, el hecho de contar con textos de diversos ámbitos y géneros discursivos proporciona un corpus representativo del español. Creemos, asimismo, que la interfaz que hemos diseñado para albergar este corpus es amigable y útil, ya que permite al usuario la selección de un subcorpus de su interés y el análisis estadístico y lingüístico del mismo. Nos parece esencial, además, el hecho de que el corpus es gratuito, se encuentra en línea y es dinámico, es decir, que estará en continuo crecimiento.

5 Trabajo futuro

Somos conscientes también de que nuestro trabajo tiene ciertas limitaciones por el momento, que intentaremos solventar en un futuro. A corto plazo, tenemos varios objetivos: añadir un anotador más para el corpus de prueba y medir el acuerdo entre los 3 anotadores; consensuar la anotación de los anotadores del corpus de prueba (probablemente empleando un "juez" externo) para establecer un conjunto de textos que puedan ser considerados como un *gold standard* preliminar para el español; analizar el corpus para detectar patrones lingüísticos que permitan la detección de relaciones, con el objetivo de desarrollar el primer analizador discursivo automático para el español. Más a largo plazo, nos planteamos otros objetivos, más ambiciosos: ampliar el corpus, añadiendo textos de nuevos dominios y géneros discursivos (con textos seleccionados por nosotros y textos enviados por los usuarios); etiquetar todos los textos del corpus por 3 personas, para finalmente contar con un verdadero *gold standard* representativo para el español (aunque este objetivo depende de la financiación del proyecto).

Agradecimientos

Este trabajo ha sido parcialmente financiado por los proyectos RICOTERM (FFI2010-21365-C03-01) y APLE (FFI2009-12188-C05-01) de España, y el proyecto 82050 del Consejo Nacional de Ciencia y Tecnología de México.

Bibliografía

Bouayad-Agha, N., L. Wanner y D. Nicklass. 2006. Discourse structuring of dynamic content. *Procesamiento del lenguaje natural*, 37:207-213.

Cabré, M.T. (1999). *La terminología: representación y comunicación*. Barcelona: IULA-UPF.

Carlson, L. y D. Marcu. 2001. Discourse Tagging Reference Manual. ISI Technical Report ISITR-545. Los Ángeles: University of Southern California.

Carlson, L., D. Marcu y M.E. Okurowski. 2002a. *RST Discourse Treebank*. Pennsylvania: Linguistic Data Consortium.

da Cunha, I., E. SanJuan, J.M. Torres-Moreno, M. Lloberes e I. Castellón. 2010. DiSeg: Un segmentador discursivo automático para el español. *Procec. del Lenguaje Natural*, 45:145-152.

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

da Cunha, I., L. Wanner y M.T. Cabré. 2007. Summarization of specialized discourse: The case of medical articles in Spanish. *Terminology*, 13(2):249-286.

Dligach, D., R.D. Nielsen y M. Palmer. 2010. To Annotate More Accurately or to Annotate More. En *Proceedings of the 4th Linguistic Annotation Workshop* (LAW-IV). 48th Annual Meeting ACL.

Hovy, E. 2010. *Annotation. A Tutorial*. Presented at the 48th Annual Meeting ACL.

Ide, N. y J. Pustejovsky. 2010. What Does Interoperability Mean, anyway? Toward an Operational Definition of Interoperability. En *Proceedings of the Second International Conference on Global Interoperability for Language Resources* (ICGL 2010).

Mann, W.C. y S.A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243-281.

Marcu, M. 2000. *The Theory and Practice of Discourse Parsing Summarization*. Massachusetts: Institute of Technology.

Marcus, M.P., B. Santorini y M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treenbank. *Computational Linguistics*, 19(2):313-330.

O'Donnell, M. 2000. RSTTOOL 2.4 – A markup tool for rhetorical structure theory.

En *Proceedings of the International Natural Language Generation Conference*, páginas 253-256.

Pradhan, S., E. Hovy, M. Marcus, M. Palmer, L. Ramshaw y R. Weischedel. 2007. OntoNotes: A Unified Relational Semantic Representation. En *Proceedings of the First IEEE International Conference on Semantic Computing* (ICSC-07).

Reitter, D. y M. Stede. 2003. Step by step: underspecified markup in incremental rhetorical analysis. En *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora*.

Romera, M. 2004. Discourse Functional Units: The Expression of Coherence Relations in Spoken Spanish. Munich: LINCOM.

Pardo, T.A.S., M.G.V. Nunes y L.H.M. Rino. 2008. DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. *Lecture Notes in Artificial Intelligence*, 3171:224-234.

Pardo, T.A.S. y E.R.M. Seno. 2005. Rhetalho: um corpus de referência anotado retoricamente. En *Anais do V Encontro de Corpora. São Carlos-SP*, Brasil.

Sierra, G. 2008. Diseño de corpus textuales para fines lingüísticos. En *Actas del IX Encuentro Internacional de Lingüística en el Noroeste*, páginas 445-462.

Stede, M. 2004. The Potsdam commentary corpus. En *Proceedings of the Workshop on Discourse Annotation, 42nd Meeting of the Association for Computational Linguistics*.

Taboada, M. 2004. *Building Coherence and Cohesion: Task-Oriented Dialogue in English and Spanish*. Amsterdam/Philadelphia: John Benjamins.

Taboada, M. y J. Renkema. 2008. *Discourse Relations Reference Corpus* [Corpus]. Simon Fraser University and Tilburg University. http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html.

Taboada, M. y W.C. Mann. 2006. Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4):567-588.

A Anexo 1: Relaciones retóricas y porcentaje de aparición en el corpus

Relación	**Tipo**	**Cantidad**	
		Nº	**%**
Elaboración	N-S	765	24,56
Preparación	N-S	475	15,25
Fondo	N-S	204	6,55
Resultado	N-S	193	6,20
Medio	N-S	175	5,62
Lista	N-N	172	5,52
Unión	N-N	160	5,14
Circunstancia	N-S	140	4,49
Propósito	N-S	122	3,92
Interpretación	N-S	88	2,83
Antítesis	N-S	80	2,57
Causa	N-S	77	2,47
Secuencia	N-N	74	2,38
Evidencia	N-S	59	1,89
Contraste	N-N	58	1,86
Condición	N-S	53	1,70
Concession	N-S	50	1,61
Justificación	N-S	39	1,25
Solución	N-S	32	1,03
Motivación	N-S	28	0,90
Reformulación	N-S	22	0,71
Conjunción	N-N	11	0,35
Evaluación	N-S	11	0,35
Disyunción	N-N	9	0,29
Resumen	N-S	8	0,26
Capacitación	N-S	5	0,16
Alternativa	N-S	3	0,10
Unless	N-S	2	0,06