La influencia del anotador y las técnicas de traducción en el desarrollo de árboles retóricos. Un estudio en español y euskera

Iria da Cunha Fanego (iria.dacunha@upf.edu)
Mikel Iruskieta Quintian (mikel.iruskieta@ehu.es)

Universitat Pompeu Fabra (Barcelona)
Universidad del País Vasco (San Sebastián)

1. Introducción

En la comunidad científica es habitual escribir los resúmenes de los artículos de investigación, además de en la lengua franca (inglés, francés…), en lenguas administrativas (portugués, español, euskera…). De hecho, en algunas revistas científicas se ha convertido en un requisito para la publicación de los artículos. Gracias a esto es posible estudiar sobre un corpus bilingüe cómo se generan en cada lengua las estructuras retóricas de los resúmenes y cómo afectan las técnicas de traducción en la estructura del discurso. Para poder realizar una evaluación adecuada del grado de acuerdo entre diferentes anotadores de estructuras retóricas, estas técnicas de traducción deberían ser identificadas. Existen ya trabajos en los que se han tratado precisamente cuestiones relacionadas con la evaluación de la anotación de la estructura retórica (Marcu et al. 1999, Marcu 2000a, Carlson et al. 2001). Sin embargo, no se ha abordado en profundidad cómo afectan estas técnicas en el proceso de anotación retórica y en la evaluación sobre el acuerdo de los anotadores.

En nuestro trabajo partimos de la *Rhetorical Structure Theory* (RST), desarrollada por Mann y Thompson (1988), por tratarse de teoría independiente de lengua. La RST es una teoría descriptiva de organización del texto muy útil para describirlo caracterizando su estructura a partir de las relaciones que mantienen entre sí los elementos discursivos o retóricos del mismo[1] (Circunstancia, Elaboración, Motivación, Evidencia, Justificación, Causa, Propósito, Antítesis, Condición, entre otras). Se basa a su vez en una serie de premisas: funcionalidad de la jerarquía, rol comunicativo de la estructura del texto y predominancia de estructuras discursivas orientadas. Así, la RST establece un inventario de relaciones entre las unidades discursivas del texto, en las que por lo general una de las unidades es el gobernante (*núcleo*), mientras que la otra (*satélite*) aporta cierta información retórica acerca de él, siendo este el esquema estructural más frecuente entre dos unidades (casi siempre adyacentes, aunque hay excepciones). Estas relaciones se denominan relaciones *nucleares* o *núcleo-satélite*. En el caso de las relaciones que no presentan una unidad central con respecto a los propósitos del autor, la relación se denomina *multinuclear*.

Esta teoría es empleada como base para investigar sobre diversos temas, mencionados en Taboada y Mann (2005), tanto teóricos como prácticos, como por ejemplo generación automática de textos, resumen automático, análisis de corpus, análisis

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

textual, traducción automática, enseñanza de redacción, adquisición de conocimiento discursivo, análisis del habla, recuperación de información, etc. Véanse, entre otros, los trabajos de Bouayad-Agha (2000), Marcu (2000a), Ghorbel et al. (2001), Ghorbel et al. (2001), Burstein y Marcu (2003), Haouam y Marir (2003), da Cunha et al. (2007) y da Cunha (2008). También se han desarrollado algunos analizadores retóricos automáticos en diversas lenguas inspirados en esta teoría: Sumita et al. (1992) en japonés, Marcu (1998) en inglés, y Pardo et al. (2004) y Pardo y Nunes (2008) en portugués de Brasil.

Pero aunque la RST ha sido ampliamente utilizada, también ha recibido algunas críticas. Stede (2008: 329), por ejemplo, critica su ambigüedad, ya que muchas de las asunciones que los anotadores realizan no se pueden hacer explícitas en un único árbol. Una evidencia de la subjetividad es, por ejemplo, lo difícil que resulta que varias personas realicen exactamente el mismo árbol retórico a partir de un mismo texto.

> An RST-style analysis of a text, on the other hand, cuts "vertically": It tries to capture the essence of coherence within a single representation structure, making a series of quite different simplifications along the way. We do not doubt that this can be an insightful instrument for studying text—RST has been quite successful for a variety of purposes. But there are inherent limitations on the explanatory power when information from different realms is conflated in a single tree structure: On the one hand, one cannot do full justice to the separate realms; on the other hand, the single tree structure becomes ambiguous, because when crafting it, many underlying assumptions cannot be made explicit. (Stede 2008: 329)

Todas las consideraciones realizadas hasta ahora nos llevan a formularnos diversas cuestiones interesantes:

- ¿Es posible comparar las estructuras retóricas de lenguas tan dispares como una lengua romance (el español) y una lengua no indoeuropea (el euskera) a partir de la misma teoría? ¿Tiene algún beneficio la comparación de las estructuras en dos lenguas tan dispares?
- ¿Hasta qué punto afecta la subjetividad al análisis retórico con la RST? ¿Cuál es la influencia del anotador en el proceso de etiquetaje retórico con relaciones de la RST?
- ¿Con qué método de evaluación podemos saber qué factores (las técnicas de traducción, el grado de abstracción teórica o la ambigüedad de la estructura retórica) afectan en la evaluación de la estructura retórica? ¿Y en qué medida afectan las técnicas de traducción al acuerdo de la estructura retórica en textos paralelos?

Responder a estas cuestiones es precisamente el objetivo de este trabajo. Para alcanzarlo, diseñamos un experimento en el que dos anotadores etiquetan retóricamente el mismo corpus de textos en español y en euskera, para luego comparar las dos anotaciones y observar las diferencias existentes entre ellos. En el apartado 2 detallamos la metodología empleada en el experimento. En el apartado 3 mostramos el análisis cuantitativo y cualitativo realizado, referente a las EDUs (*Elementary Discourse Units*), a la nuclearidad y a las relaciones retóricas. En el apartado 4 exponemos las conclusiones del trabajo.

2. Metodología

La metodología empleada para realizar este estudio incluye varias fases. En primer lugar, conformamos un corpus de análisis. En segundo lugar, definimos criterios de partida con respecto a la segmentación de las EDUs y a las relaciones empleadas en el trabajo. En tercer lugar, dos anotadores etiquetamos los textos del corpus (en castellano y en euskera). En cuarto lugar, realizamos el análisis cuantitativo. En quinto lugar, llevamos a cabo el análisis cualitativo. Finalmente, extraemos conclusiones de ambos análisis.

2.1. Corpus

El corpus de análisis está formado por 20 resúmenes en español y en euskera incluidos en artículos médicos de investigación extraídos de la Gaceta Médica de Bilbao[2] escritos entre los años 2000 y 2008 por diferentes especialistas. Esta revista solicita a los autores los resúmenes de sus artículos en español, euskera e inglés. El hecho de que sea el mismo autor quien redacte los resúmenes en las tres lenguas es una garantía de que dichos resúmenes incluyen la misma información y de que siguen una estructura similar.

2.2. Criterios de partida

Antes de llevar a cabo el análisis retórico de los textos del corpus, establecemos algunos criterios en relación con la segmentación de las EDUs y las relaciones retóricas empleadas en el estudio.

2.2.1. Segmentación de las EDUs

Con respecto a la segmentación de las EDUs, revisamos algunas cuestiones del manual de Carlson y Marcu (2001), del cual partimos para realizar la segmentación inicial. Estas puntualizaciones se realizan partiendo de la idea de que queremos diferenciar claramente los niveles sintáctico y discursivo. En nuestro trabajo consideramos que, en principio, las EDUs deben incluir un verbo (es decir, constituir una oración o una cláusula) y reflejar realmente una relación retórica. Dichas puntualizaciones son las siguientes:[3]

a) En Carlson y Marcu (2001) los complementos de verbos de atribución (actos de habla u otros actos cognitivos) son tratados como EDUs, como en el ejemplo 1a:[4]

1a. [Bush indicated] [there might be “room for flexibility” in a bill] [..]

En cambio, nosotros no trataremos estos complementos de verbos de atribución como EDUs, y segmentaríamos el mismo fragmento como se muestra en el ejemplo 1b:

1b. [Bush indicated there might be “room for flexibility” in a bill] [..]

---

[2] http://www.gacetamedicabilbao.org/web/es/

[3] Los ejemplos mostrados a continuación se han extraído de Carlson y Marcu (2001).

[4] Los ejemplos marcados con una “a” reflejan la segmentación realizada en Carlson y Marcu (2001) y los ejemplos marcados con una “b” muestran la segmentación que realizaremos en nuestro trabajo.

La cláusula “there might be ‘room for flexibility’ in a bill” constituye un objeto directo (desde el punto de visto de la sintaxis de constituyentes) o un actante II (desde la sintaxis de dependencias) del verbo “indicate” y, por este motivo, en nuestro trabajo lo consideraremos solo en dicho plano.

b) En Carlson y Marcu (2001) se especifica que en el ejemplo 2a las cláusulas que dependen de “so that their clients can” son segmentadas como EDUs separadas, las cuales a su vez son consideradas como satélites de una relación de Propósito. El satélite consistiría en una Lista multinuclear de cláusulas coordinadas:

> 2a. [Equipped with cellular phones, laptop computers, calculators and a pack of blank checks,] [they parcel out money] [so that their clients can find temporary living quarters,] [buy food,] [replace lost clothing,] [repair broken water heaters,] [and replaster walls.]

Por el contrario, nosotros trataríamos todas estas cláusulas como una única EDU:

> 2b. [Equipped with cellular phones, laptop computers, calculators and a pack of blank checks,] [they parcel out money] [so that their clients can find temporary living quarters, buy food, replace lost clothing, repair broken water heaters, and replaster walls.]

Como en el caso anterior, consideramos que todas estas cláusulas constituyen un objeto directo o un actante II del verbo “can”.

3. En Carlson y Marcu (2001) las cláusulas de relativo, las cláusulas que modifican elementos nominales o las cláusulas que truncan otras EDUs son tratadas como unidades discursivas “embedded”, mientras que nosotros no las consideraremos como tal. Veamos caso por caso.

- Cláusulas de relativo:

> 3a. [A separate inquiry by Chemical cleared Mr. Edelson of allegations] [*that* he had been lavishly entertained by a New York money broker.]
> 3b. [A separate inquiry by Chemical cleared Mr. Edelson of allegations *that* he had been lavishly entertained by a New York money broker.]

- Cláusulas que modifican elementos nominales:

> 4a. [The results underscore Sears’s difficulties] [*in implementing* the “everyday low pricing” strategy] [that it adopted in March, as part of a broad attempt] [*to revive* its retailing business.]
> 4b. [The results underscore Sears’s difficulties *in implementing t*he “everyday low pricing” strategy that it adopted in March, as part of a broad attempt *to revive* its retailing business.]

- Aposiciones:

> 5a. [The fact] [*that* this happened two years ago] [and there was a recovery] [gives people some comfort] [*that* this won’t be a problem.]
> 5b. [The fact *that* this happened two years ago and there was a recovery gives people some comfort *that* this won’t be a problem.]

- Elementos parentéticos:

> 6a. [The Tass news agency said the 1990 budget anticipates income of 429.9 billion rubles] [*($US693.4 billion)*] [and expenditures of 489.9 billion rubles] [*($US790.2 billion)*.]
> 6b. [The Tass news agency said the 1990 budget anticipates income of 429.9 billion rubles *($US693.4 billion)* and expenditures of 489.9 billion rubles *($US790.2 billion)*.]

En nuestro trabajo solo segmentaremos los fragmentos entre paréntesis si estos realmente conforman una EDU, es decir, un elemento que mantiene alguna relación discursiva con otro elemento y que contiene un verbo.

- Cláusulas coordinadas en unidades "embedded":

> 7a. [She signed up,] [starting as an "inside" adjuster,] [who settles minor claims] [*and* does a lot of work by phone.]
> 7b. [She signed up,] [starting as an "inside" adjuster, who settles minor claims *and* does a lot of work by phone.]

4. En Carlson y Marcu (2001) las frases que comienzan con un marcador discursivo muy evidente, como "porque", "a pesar de", "como resultado de", etc. son tratadas como EDUs, como puede observarse en los ejemplos 8a y 9a:

> 8a. [But some big brokerage firms said] [they don't expect major problems] [*as a result of* margin calls.]
> 9a. [Today, no one gets in or out of the restricted area] [*without* De Beers's stingy approval.]

En nuestro trabajo solo consideraremos como EDUs las frases que empiecen con estos marcadores pero contengan además un verbo. Por tanto, nosotros segmentaríamos los ejemplos anteriores de la siguiente manera:

> 8b. [But some big brokerage firms said they don't expect major problem *as a result of* margin calls.]
> 9b. [Today, no one gets in or out of the restricted area *without* De Beers's stingy approval.]

5. En Carlson y Marcu (2001) se establecen una serie de convenciones sobre puntuación de cara a la segmentación. En nuestro trabajo no emplearemos algunas de estas convenciones como marcas para la segmentación, a no ser que los segmentos marcados contengan un verbo y sí sean en realidad un elemento discursivo. Veamos algunos ejemplos:

- Paréntesis:

> 10a. [If the government can stick with them,] [it will be able to halve this year's 120 billion ruble] [*(US$193 billion)*] [deficit.]
> 10b. [If the government can stick with them,] [it will be able to halve this year's 120 billion ruble *(US$193 billion)* deficit.]

- Guiones:

> 11a. [This will require us to define] [- *and redefine* -] [what is 'necessary' or 'appropriate' care.]
> 11b. [This will require us to define - *and redefine* - what is 'necessary' or 'appropriate' care.]

Con respecto a la utilización de otros signos de puntuación (coma, punto, punto y coma, etc.) como marca para la segmentación, estamos de acuerdo con Carlson y Marcu (2001: 30):

> Commas and periods are not independent justification for an EDU boundary. If a unit is a legitimate EDU and it ends with a comma or period, the punctuation is included as part of that EDU.

Finalmente, es importante recalcar que, al igual que ocurre en Carlson y Marcu (2001), si observamos que una EDU es truncada por otra (es decir, incluye otra EDU), los dos fragmentos de esa primera EDU se segmentarán y posteriormente se marcarán con una relación de *Same-unit*. Así, el ejemplo 12 se etiquetará como refleja la Figura 1:

> 12. [Las válvulas ahorradoras de oxígeno (VAO),] [al liberar oxígeno únicamente durante la inspiración,] [evitan que se pierda durante la fase respiratoria,] […]

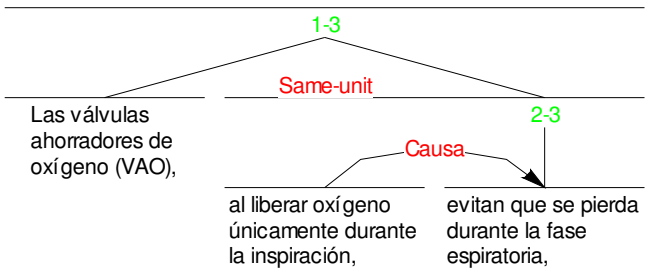

Figura 1. Árbol retórico que muestra la relación de *Same-unit*

2.2.2. Relaciones retóricas

En lo que se refiere a la detección de las relaciones retóricas y la nuclearidad (es decir, en cuanto a la decisión de considerar un segmento como núcleo o como satélite):

a) Se consensúa la lista de relaciones retóricas de la RST. Dentro de la RST existen diferentes clasificaciones de las relaciones retóricas: la clásica de Mann y Thompson de 24 relaciones (Mann y Thompson 1988), la extendida de Mann y Thompson de 30 relaciones (Mann 2005) y la de Marcu de 136 relaciones (Carlson et al. 2001) entre otras. Hemos elegido la clasificación extendida para la anotación del corpus paralelo por ser la que mejor se adecua a nuestros propósitos de segmentación. De todas maneras, tal como señalan Marcu et al. (1999: 55), la reducción de la taxonomía de relaciones no tiene un impacto significativo sobre el acuerdo entre los analistas:

> The results […] show that a significant reduction in the size of the taxonomy of relations may not have a significant impact on agreement ($k_{\gamma\gamma}$ is only about 4% higher than k_{γ}). This suggests that choosing one relation from a set of rhetorically similar relations produces some, but not too much, confusion.

b) Se busca un ejemplo real representativo de cada relación y se marcan los núcleos y/o los satélites en cada uno. Los ejemplos se extraen del corpus empleado en da Cunha (2008), formado también por artículos médicos en español de la revista Medicina Clínica[5]. Una vez seleccionados los ejemplos en español, estos se traducen al euskera y se marcan sus núcleos y sus satélites.

2.3. Anotación retórica

Una vez establecidos los criterios de partida, los dos anotadores etiquetamos retóricamente los 20 textos del corpus (uno en español [A1] y otro en euskera [A2]), empleando las relaciones de la RST. La anotación se divide en dos fases principales: segmentación de las EDUs y análisis retórico.

2.3.1. Segmentación de las EDUs

En esta fase, cada uno de los anotadores segmenta las EDUs de los 20 resúmenes, por separado y sin consultarse entre ellos. Para segmentar las EDUs se emplea la herramienta RSTTool (O'Donnell, 2000).[6]

Una vez recogidos los datos sobre el acuerdo entre las segmentaciones realizadas por los dos anotadores, realizamos un pequeño debate para igualar la segmentación de los resúmenes en español y en euskera. Esta homogeneización de los segmentos en ambos tipos de resúmenes se realiza con la intención de minimizar el ruido que pueda surgir de una segmentación diferente y así proceder a una evaluación más fina sobre la nuclearidad y las relaciones en los árboles retóricos. Esta comparación, como veremos, se realiza de manera manual (evaluando precisión y cobertura), debido a la carencia actual de herramientas automáticas de comparación de árboles retóricos en diferentes lenguas. En el Núcleo Interinstitucional de Lingüística Computacional (NILC) se ha desarrollado la herramienta RSTc Tool for Discourse Parsing Evaluation[7], que compara árboles retóricos, pero en una misma lengua, por tanto no podemos emplearla en este estudio. Como nuestra comparación ha de ser entonces manual, consideramos oportuno realizar en nuestro trabajo esta homogeneización de las EDUs, para que los anotadores partan de los mismos segmentos, puedan establecer relaciones entre ellos, construir los árboles retóricos y realizar finalmente la comparación entre estos de una manera mucho más sencilla.

2.3.2. Análisis retórico

En esta fase cada anotador etiqueta retóricamente la segmentación homogeneizada de los 20 resúmenes por separado, marcando las relaciones entre las EDUs y determinando cuáles de estas EDUs son núcleos y cuáles satélites. Para ello se emplea de nuevo la RSTTool y la clasificación extendida de relaciones retóricas.

2.4. Análisis cuantitativo

Después de haber realizado la anotación, realizamos un análisis cuantitativo de los dos aspectos mencionados en el apartado anterior.

---

[5] http://dialnet.unirioja.es/servlet/revista?tipo_busqueda=CODIGO&clave_revista=2426
[6] http://www.wagsoft.com/RSTTool/
[7] http://www.nilc.icmc.usp.br/~erick/rstc2/

2.4.1. Segmentación de las EDUs

El contraste entre las segmentaciones de las EDUs realizadas por los dos anotadores se lleva a cabo mediante la evaluación de la precisión y la cobertura. Para medir la precisión, se observa cuántas EDUs seleccionadas por el A2 coinciden con las EDUs seleccionadas por el A1. Para medir la cobertura se contabiliza el número total de EDUs detectadas por el A2, en relación con el número total de EDUs detectadas por el A1. Este análisis se realiza, por un lado, sobre cada texto individualmente y, por otro, sobre el conjunto de textos del corpus.

2.4.2. Análisis retórico

Para cuantificar las coincidencias en el análisis retórico realizado por los dos anotadores hemos seguido el método de Marcu (2000b). Así, obtenemos datos referentes a los nodos (es decir, los conjuntos de EDUs relacionados retóricamente)[8], a la nuclearidad y a las relaciones retóricas detectadas.

Para comparar los análisis retóricos realizados por los dos anotadores se miden de nuevo la precisión y la cobertura. Para medir la precisión, se observa cuántos nodos, núcleos y satélites, y relaciones retóricas seleccionados por el A2 coinciden con los seleccionados por el A1. Para medir la cobertura se contabiliza el número total de los mismos elementos detectados por el A2, en relación con el número total detectado por el A1. De nuevo, este análisis se realiza, por un lado, sobre cada texto individualmente y, por otro, sobre el conjunto de textos del corpus. Veamos un ejemplo. La Figura 2 muestra un fragmento de árbol retórico en español realizado por el A1, mientras que la Figura 3 muestra el árbol retórico del mismo texto en euskera, realizado por el A2.

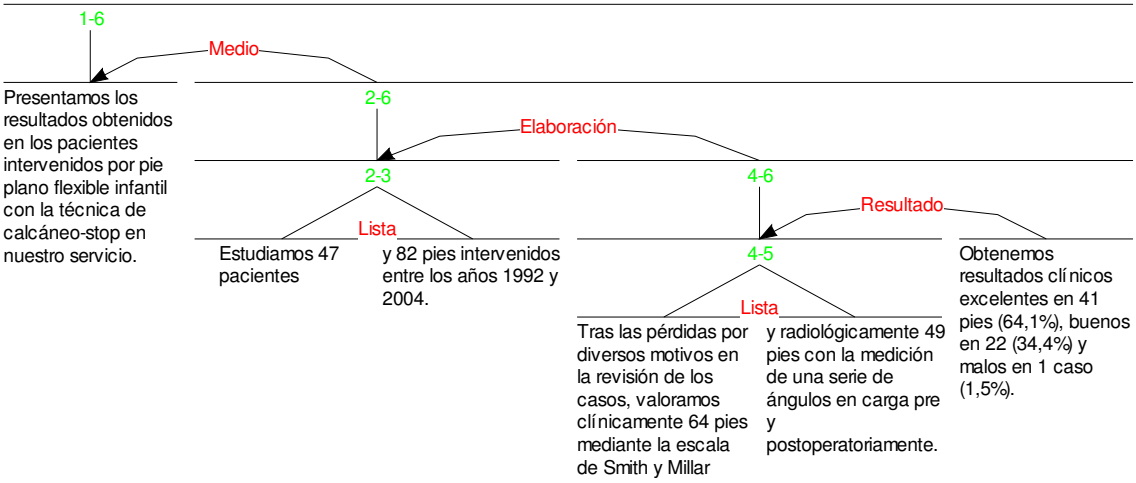

Figura 2. Árbol retórico en español realizado por el A1

[8] Marcu (2000b) los denomina “spans”.

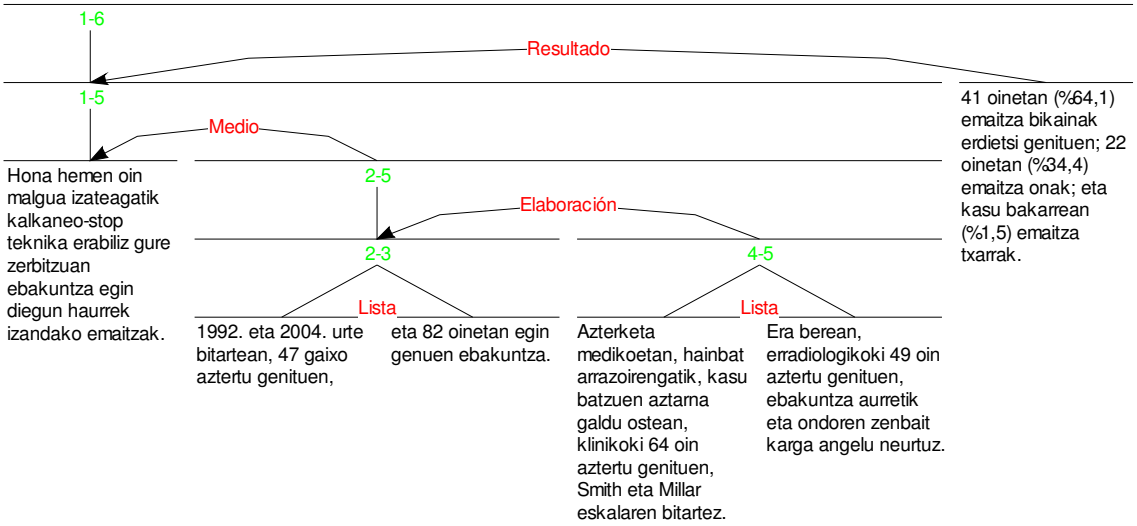

Figura 3. Árbol retórico en euskera realizado por el A2

La Tabla 1 refleja la metodología de la evaluación de Marcu (2000b): se indican las EDUs, los nodos, los núcleos (N) y satélites (S), y las relaciones detectadas tanto por el A1 como por el A2. A la hora de incluir el tipo de relación en la tabla, los núcleos se marcan como NÚCLEO[9] (excepto en las relaciones multinucleares) y los satélites con el nombre de la relación retórica (por ejemplo, Resultado, Elaboración, Medio, etc.). Ha de tenerse en cuenta que, en nuestro trabajo, al haber homogeneizado las EDUs en la fase de segmentación (véase apartado 2.3.1.), las EDUs detectadas por el A1 y el A2 siempre coincidirán. En la Tabla 1, hemos marcado en gris las diferencias entre los dos anotadores.

	EDU	**EDU**	**Nodo**	**Nodo**	**Nucle-aridad**	**Nucle-aridad**	**Relación**	**Relación**
Consti-tuyente	**A1**	**A2**	**A1**	**A2**	**A1**	**A2**	**A1**	**A2**
1-1	X	X	X	X	N	N	NÚCLEO	NÚCLEO
2-2	X	X	X	X	N	N	LISTA	LISTA
3-3	X	X	X	X	N	N	LISTA	LISTA
4-4	X	X	X	X	N	N	LISTA	LISTA
5-5	X	X	X	X	N	N	LISTA	LISTA
6-6	X	X	X	X	S	S	RESULTADO	RESULTADO
4-5			X	X	N	S	NÚCLEO	ELABORACIÓN
4-6			X	-	S	-	ELABORACIÓN	-
2-3			X	X	N	N	NÚCLEO	NÚCLEO
2-6			X	-	S	-	MEDIO	
2-5			-	X	-	S	-	MEDIO
1-5			-	X	-	N	-	NÚCLEO

Tabla 1. Evaluación cuantitativa según el método de Marcu (2000b)

Una vez realizada esta tabla se mide la precisión y la cobertura de la manera detallada más arriba. La Tabla 2 muestra los resultados de esta evaluación. Observamos en este ejemplo que la cobertura es del 100% en todos los casos, mientras que la precisión oscila entre un 80% (nodos) y un 70% (nuclearidad y relaciones retóricas).

[9] Marcu (2000b) los denomina “spans”.

	Total cobertura	Total precisión
Nodos	100%	80%
Nuclearidad	100%	70%
Relaciones	100%	70%

Tabla 2. Resultados de la evaluación cuantitativa de los árboles retóricos de las Figuras 2 y 3

2.5. Análisis cualitativo

En el análisis cualitativo nos centramos también en cuestiones referentes a la segmentación de las EDUs y al análisis retórico.

2.5.1. Segmentación de las EDUs

Después de haber cuantificado las diferencias en la segmentación de las EDUs realizada por los dos anotadores, se observan los casos concretos en los que estos difieren y se estudian las posibles razones del desacuerdo.

Al igualar u homogenizar las EDUs, observamos que algunas de ellas entran en contradicción con las pautas de segmentación establecidas. Esto es debido a que las técnicas de traducción afectan también a la segmentación. Por ejemplo, en español hay fragmentos que se consideran como una única EDU, pero que se segmentan en dos para llevar a cabo la homogeneización:

12a. [Se realiza el estudio de la proteína 14-3-3, que resulta ser positivo.]

12b. [14-3-3 proteinaren azterketa egin zaio,] [eta emaitza positiboak lortu dira.]

En el ejemplo 12a se observa que el anotador ha segmentado este fragmento en español como una sola EDU, ya que las oraciones de relativo no se consideran como tal. En cambio, el ejemplo 12b refleja que en euskera la oración de relativo se ha traducido en una oración principal, que se relaciona con la oración anterior con un marcador del discurso de conjunción ("eta"). Para homogeneizar decidimos segmentar la EDU en español en dos EDUs, de la manera que se muestra en el ejemplo 12c:

12c. [Se realiza el estudio de la proteína 14-3-3,] [que resulta ser positivo.]

2.5.2. Análisis retórico

El método de evaluación de Marcu (2000b), ejemplificado en el apartado 2.4.2., nos parece válido. No obstante, este método solo considera el acuerdo absoluto en todos los niveles. Así, un desacuerdo a nivel de segmentación o un desacuerdo en los nodos inferiores afectará de un modo definitivo en el acuerdo sobre las relaciones retóricas superiores en el árbol. Por ejemplo, en los árboles retóricos de las Figuras 2 y 3, con el método de Marcu (2000b) observamos desacuerdo en la detección de nodos, nuclearidad y relaciones. Sin embargo, las cinco relaciones marcadas por los dos anotadores coinciden. Habría entonces diferencias en cuanto a los nodos detectados, pero no en cuanto a las relaciones. Consideramos que es necesario realizar también este tipo de aproximación, en cierta medida más optimista, a la que llamaremos "evaluación parcial cualitativa". En las Tablas 3 y 4 se incluyen los datos de esta evaluación, referidos a los nodos y a las relaciones, respectivamente.

Constituyentes		Nodos	
A1	**A2**	**A1**	**A2**
4-5	4-5	X	X
2-3	2-3	X	X
2-6	2-5	X	X
1-6	1-5	X	X
4-6	1-6	-	-

Tabla 3. Evaluación parcial cualitativa de los nodos

Relaciones detectadas	
A1	**A2**
Lista	Lista
Lista	Lista
Elaboración	Elaboración
Medio	Medio
Resultado	Resultado

Tabla 4. Evaluación parcial cualitativa de relaciones retóricas

En la Tabla 5 se muestran los resultados de la evaluación parcial cualitativa de este ejemplo. Observamos que la cobertura y la precisión es del 100% en todos los casos, excepto la precisión de los nodos, que obtiene un 80%.

	Total cobertura	**Total precisión**
Nodos	100%	80%
Nuclearidad	100%	100%
Relaciones	100%	100%

Tabla 5. Resultados de la evaluación parcial cualitativa de los árboles retóricos de las Figuras 2 y 3

Ya que obtendremos resultados cuantitativos con el método de Marcu (2000b) y considerando que nos enfrentamos a la anotación de textos en diferentes lenguas, nos centraremos en la evaluación parcial cualitativa únicamente de las relaciones retóricas, y no de los nodos y de la nuclearidad.

En la evaluación parcial cualitativa analizaremos de una manera sistemática las causas de las discrepancias entre anotadores. Por un lado, observaremos los fenómenos que pueden crear diferencias en el acuerdo sobre la anotación que mencionan Mann y Thompson (1987): ambigüedad en la estructura del texto, análisis simultáneos y errores analíticos, entre otros. Por otro lado, analizaremos el fenómeno que se refleja en Marcu et al. (2000: 10), consistente en cambiar el tipo de relación retórica al traducir de una lengua a otra:

> Hence, the mappings in (4) provide an explicit representation of the way information is re-ordered and re-packaged when translated from Japanese into English. However, when translating text, it is also the case that the rhetorical rendering changes. What is realized in Japanese using an CONTRAST relation can be realized in English using, for example, a COMPARISON or a CONCESSION relation.

De este modo trataremos de detectar las posibles causas de las discrepancias entre anotadores y la influencia que tienen las técnicas de traducción en la estructura retórica, que se exponen en el apartado 3.2.

Para el cómputo de la cantidad total de las relaciones ha de tenerse en cuenta que cada relación nuclear se cuenta como una relación, mientras que las relaciones multinucleares se cuentan como funciones binarias. Por ejemplo, una relación de Lista con cuatro núcleos se representa uniendo los núcleos de manera binaria, obteniendo así tres relaciones multinucleares, cada una con dos núcleos. En las Figuras 4 y 5 se observa la anotación tradicional de la relación de Lista y la anotación binaria, respectivamente.

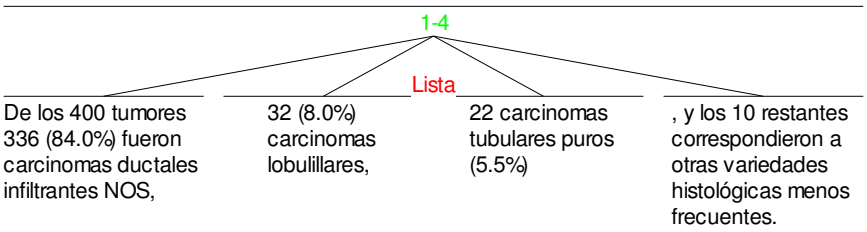

Figura 4. Anotación tradicional de una relación multinuclear de Lista

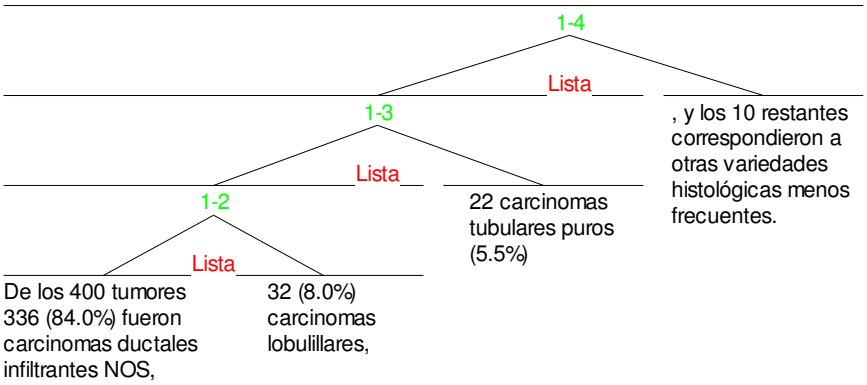

Figura 5. Interpretación binaria de una relación multinuclear de Lista

De esta forma, además de contar de un modo adecuado las relaciones multinucleares, podría compararse, por ejemplo, una confusión en la que toman parte tres unidades o nodos de una relación de Lista con tres núcleos (del A1), con una relación de Lista con dos núcleos más una relación de Elaboración (del A2). Si no se hiciese de este modo, no podríamos comparar una única relación de Lista del A1 con una única relación de Lista y una relación de Elaboración. Además, no sería adecuado contar todas las unidades nucleares de la relación de Lista porque si no las relaciones multinucleares tendrían más peso en la evaluación cualitativa.

3. Resultados

En este apartado se exponen los resultados del análisis cuantitativo y cualitativo del experimento.

3.1. Análisis cuantitativo

3.1.1. Segmentación de las EDUs

El número de EDUs segmentadas por el A1 en los textos en español es de 206, mientras que el número de EDUs segmentadas por el A2 en los textos en euskera es de 238. Los dos anotadores coinciden al segmentar 152 EDUs iguales. Siguiendo la metodología expuesta en el apartado 2.4.1., obtenemos la precisión y la cobertura de la segmentación realizada: un 63,9% y un 86,6%, respectivamente. Los casos en que la segmentación resultó diferente estuvieron motivados en su mayoría por la influencia de las técnicas de traducción (85 casos), que veremos con detalle en el apartado 3.2.1.

3.1.2. Análisis retórico

Del modo expuesto en el apartado 2.4.2., medimos la precisión y la cobertura para evaluar la coincidencia entre los nodos, la nuclearidad y las relaciones detectados por los dos anotadores. En la Tabla 6 se muestran los resultados finales. Se observa que los resultados referentes a la cobertura son similares, lo cual es debido a la homogeneización de las EDUs comentada en el apartado 2.3.1. En cambio, los resultados referentes a la precisión varían. Notamos que la precisión alcanzada es muy elevada en todos los casos. La coincidencia entre los nodos detectados por los dos anotadores es de un 92,5%, la coincidencia con respecto a la nuclearidad supone un 82,1% y la coincidencia en cuanto a las relaciones detectadas es de un 73,4%.

	Total cobertura	**Total precisión**
Nodos	98.6%	92.5%
Nuclearidad	98.6%	82.1%
Relaciones	98.6%	73.4%

Tabla 6. Resultados de la evaluación cuantitativa del estudio

3.2. Análisis cualitativo

3.2.1. Segmentación de las EDUs

Notamos que, en ocasiones, las diferencias lingüísticas entre los textos en euskera y en español provocan que los anotadores segmenten de manera diferente un mismo fragmento (véase ejemplo 14).

> 14a. [Hemos estudiado retrospectivamente 23 infecciones protésicas de rodilla tratadas en nuestro hospital entre el año 1996 y el 2004 de las cuales hemos excluido 6 por diferentes motivos.]
>
> 14b. [1996. eta 2004. urteen bitartean gure ospitalean izandako 23 infekzio protesiko aztertu ditugu.] [Horien artean, 6 kasu baztertu ditugu hainbat arrazoiengatik.]

En el ejemplo 14a se observa que el A1 ha segmentado el fragmento en español en una única unidad discursiva. En cambio, en el ejemplo 14b, vemos que el A2 ha segmentado el mismo fragmento en euskera en dos unidades. Esta disconformidad en la fase de segmentación deja patente, por un lado, que la oración de relativo no se considera como

una EDU y, por otro, que la estructura sintáctica de relativo se ha traducido al euskera en otra oración separada por un signo de puntuación muy evidente, el punto.

Al realizar la evaluación de la segmentación nos encontramos con la adversidad ya mencionada por Carson y Marcu (2001: 2), quienes afirman que la frontera entre el discurso y la sintaxis puede ser en ocasiones muy resbaladiza:

> The first step in characterizing the discourse structure of a text in our protocol is to determine the elementary discourse units (EDUs), which are the minimal building blocks of a discourse tree. Mann and Thompson (1988, p. 244) state that "RST provides a general way to describe the relations among clauses in a text, whether or not they are grammatically or lexically signalled." Yet, applying this intuitive notion to the task of producing a large, consistently annotated corpus is extremely difficult, because the boundary between discourse and syntax can be very blurry.

Las técnicas de traducción empleadas, precisamente, son una de las causas que influyen en las decisiones de segmentación. Veamos el ejemplo 15:

> 15a. [Se han estudiado un total de 442 cánceres de mama unifocales de 2 cm o menos en la pieza histológica (pT1) *operados* entre enero de 1993 y diciembre de 2005.]
>
> 15b. [Guztira, foku bakarreko 442 bularreko minbizi aztertu dira, pieza histologikoan (pT1) 2 cm edo gutxiago dituztenak.] [Guztiak 1993ko urtarrilaren eta 2005eko abenduaren artean *operatu ziren*.]
>
> 15c. [Se han estudiado un total de 442 cánceres de mama unifocales de 2 cm o menos en la pieza histológica (pT1).] [Todos *fueron operados* entre enero de 1993 y diciembre de 2005.] [Traducción del ej. 15b]

En este caso, se ha decidido traducir una forma no personal del verbo (un participio, "operado") en una forma personal ("fueron operados"), y además separar las dos oraciones mediante un punto, lo que afecta de manera contundente a la segmentación en una y otra lengua.

Observamos diversas técnicas de traducción que afectan a la segmentación realizada por los dos anotadores. En concreto, para traducir al euskera se han utilizado especialmente dos técnicas, que suponen el 74,28% de todas las técnicas de traducción y que afectan en el nivel de segmentación:

- Las cláusulas subordinadas de relativo en español se han traducido como oraciones separadas en euskera.
- En euskera se recuperan los elementos omitidos de las elipsis y de las anáforas en español y con ellos se constituyen nuevas oraciones.

Las consecuencias derivadas del empleo de estas técnicas de traducción son las siguientes:

- En euskera se utilizan más EDUs que en español. En concreto, en nuestro corpus, hay un 13,45% de EDUs más en euskera.
- Esta diferencia entre las EDUs en ambas lenguas afecta de forma considerable al acuerdo en la segmentación, y por lo tanto de una manera gradual en los demás factores a evaluar (nodos, nuclearidad y relaciones), lo que dificulta tanto la evaluación cuantitativa como la cualitativa.

Es de destacar que ha habido un acuerdo casi total en la segmentación de las EDUs en las que no ha influido ninguna técnica de traducción y que los errores de segmentación de los anotadores han sido mínimos.

3.2.2. Análisis retórico

Principalmente se han observado dos tipos de situaciones:

1. Ambigüedad o interpretaciones dispares en la elección de las relaciones: los anotadores etiquetan de forma diferente algunas relaciones que pueden resultar ambiguas. Por ejemplo, mientras que el A1 selecciona la relación de Fondo, el A2 escoge para el mismo fragmento la relación de Elaboración (véase ejemplo 16).

> 16a. [Han participado 92 pacientes ingresados en un Área Médica del Hospital de Basurto (Bilbao).]N [Todos los pacientes fueron entrevistados para elaborar la historia patopsicobiográfica necesaria para aplicar la Clasificación Psicosomática de Pierre Marty.]S_Elaboración

> 16b. [Basurtoko (Bilbo) Ospitaleko Medikuntza Arlo batean ospitaleratuta dauden 92 gaixok parte hartu dute.]S_Fondo [Pierre Martyren Sailkapen Psikosomatikoa aplikatzeko beharrezkoa den historia patopsikobiografikoa egiteko asmoz, elkarrizketa egin zitzaien gaixo guztiei.]N

Nótese que, en este caso, ha sido la discrepancia en torno a la nuclearidad la que ha llevado a seleccionar diferentes relaciones. En el ejemplo 16 se observa que en el fragmento en español el núcleo es la primera EDU (los participantes del estudio), mientras que en euskera el núcleo es la segunda EDU (el método seguido para el estudio).

Veamos otros ejemplos:

> 17a. [Se estima que el 80% de los usuarios acuden por iniciativa propia a los servicios de urgencia]N_Lista [y que el 70% de las consultas son consideradas leves por el personal sanitario.]N_Lista

> 17b. [Erabiltzaileen %80ak bere kabuz erabakitzen dute larrialdi zerbitzu batetara jotzea]N [eta kontsulta hauen %70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek.]S_Elaboración

En el ejemplo 17 también ha habido una discrepancia en torno a la nuclearidad. Sin embargo, en este caso la discrepancia afecta al carácter mismo de la relación, ya que el A1 ha anotado una relación paratáctica de Lista, mientras que el A2 ha anotado una relación hipotáctica de Elaboración.

> 18a. [Por lo demás existen buenos indicadores de proceso]S_Antítesis [pero se aprecia un escaso registro de la capacidad funcional del paciente al alta, que dificulta la comparación de los resultados de la atención sanitaria.]N

> 18b. [Gainerakoan, pozesu adierazle egokiak daude,]N [baina altan dagoen gaixoaren lamen funtzionalaren erregistro urria antzematen da, eta horrek osasun arretaren emaitzen alderaketa zailtzen du.]S_Concesión

En el ejemplo 18 la discrepancia se ha debido al distinto significado de la relación, ya que los dos anotadores han marcado una relación hipotáctica de presentación, pero mientras que el A1 ha anotado la relación de Antítesis, el A2 ha marcado la relación de Concesión.

2. Diferencias en cuanto a las técnicas de traducción entre el euskera y el español: las diferencias lingüísticas entre estas dos lenguas provocan que, en ocasiones, los anotadores etiqueten de manera diferente un mismo fragmento (véanse ejemplos 19 y 20).

> 19a. [Escogiendo la especialidad más barata existente en el mercado]S_Circunstancia [podríamos alcanzar un ahorro de 6.463.400,35 €.]N

> 19b. [Merkatuak eskaintzen digun espezialitate merkeena aukeratuko bagenu]S_Condición [6.463.400,35€-ko aurrezpena lortuko genuke.]N

El gerundio ("escogiendo") puede indicar en español una relación de Circunstancia. En euskera no se incluye un gerundio en la oración, si no una marca condicional en uno de los verbos, lo que evidencia una relación de Condición.

> 20a. [En los 7 ítems se han encontrado diferencias estadísticamente significativas entre el grupo de pacientes oncológicos con los pacientes afectos de otro tipo de patologías ($p<0.05$).]N [*Estos ítems* diferencian a los pacientes con neoplasias de otro tipo de pacientes, y permiten una valoración global de los mismos, ofreciendo una idea de las expectativas del proceso.]S_Elaboración

> 20b. [7 itemak aztertuta, estatistikoki desberdintasun aipagarriak aurkitu ziren gaixo onkologikoen eta bestelako patologiak dituzten gaixoen artean ($p<0.05$).]N_Unión [*Horrez gain*, item horiek neoplasiak dituzten gaixoak eta bestelako gaixoak bereizten dituzte, horiei buruzko balorazio orokorra egiteko aukera ematen dute, eta prozesuaren igurkapenen gaineko argibideak ematen dizkigute.]N_Unión

En español se ha interpretado la relación de Elaboración debido a la anáfora, mientras que en euskera el marcador discursivo "horrez gain", equivalente en español a "además de eso", introduce un nuevo tópico en el discurso, por lo que se marca una relación multinuclear. Por tanto, es evidente que una estrategia diferente en la traducción afecta al análisis retórico del texto.

Hemos analizado este fenómeno sistemáticamente e incluimos en la Tabla 7 las técnicas de traducción empleadas en español y en euskera, con sus respectivas frecuencias.

Técnicas de traducción	**Español**	**Euskera**	**Total**
a) Completar elipsis y separar con puntuación	1	5	6
b) Emplear formas personales del verbo	0	5	5
c) Usar marcadores del discurso	2	7	9
d) Suprimir oraciones de relativo	0	6	6
e) Otras técnicas	0	5	5
Total	3	28	31

Tabla 7. Técnicas de traducción que determinan diferentes relaciones retóricas

A continuación ofrecemos algunos ejemplos:

a) Completar elipsis y separar con puntuación:

> 21a. [Todos los pacientes presentaban una insuficiencia ventilatoria, en 10 casos de tipo obstructivo y en los restantes de tipo no obstructivo o mixto.]
>
> 21b. [Gaixo guztiek zukaten aireztapen gutxiegitasuna;] [hamar kasutan butxaketa-motakoa *zen*] [eta gainerakoetan ezbutxaketakoa edo mistoa *zen*.]

En este caso, se ha traducido este fragmento al euskera completando la elipsis de los verbos en la descripción de los casos de "insuficiencia ventilatoria".

b) Emplear formas personales del verbo:

> 22a. [Estudiamos 47 pacientes y 82 pies *intervenidos* entre los años 1992 y 2004.]
>
> 22b. [1992. eta 2004. urte bitartean, 47 gaixo aztertu genituen,] [eta 82 oinetan *egin genuen ebakuntza*.]

La forma impersonal del verbo en participio del español ("intervenidos") se ha traducido al euskera mediante una estructura que utiliza una forma personal del verbo y su objeto directo ("egin genuen ebakuntza").

> 23a. [Nuestros resultados sugieren la presencia de alteraciones respiratorias crónicas con el resultado de un déficit ventilatorio, varias décadas después del tratamiento con colapsoterapia; *comprobando* una buena respuesta al tratamiento con ventilación domiciliaria.]
>
> 23b. [Gure emaitzek iradokitzen dute kolapsoterapiarekin egindako tratemendutik hamarkada batzuk gerago arnas alterazio kronikoak daudela aireztapen déficit baten emaitzarekin;] [eta *egiaztatu d*a etxeko aireztapenarekin egindako tratamenduak erantzun ona izan duela.]

En este ejemplo ha sido el gerundio en español ("comprobando") el que se ha transformado en verbo personal ("egiaztatu da").

c) Usar marcadores del discurso:

> 24a. [Como cirugía primaria (…) presenta una mortalidad del 0,5%] [y un 8,8% de complicaciones perioperatorias, destacando la hemorragia (4,8%) y la dehiscencia anastomótica (1,7%).]
>
> 24b. [Kirurgia mota honetan, heriotza tasa % 0,5ekoa da,] [eta ebakuntza osteko arazoak, *berriz*, % 8,8koak dira: odoljarioa (% 4,8) eta dehiszentzia anastomotikoa (% 1,7).]

La utilización en euskera del marcador del discurso "berriz", equivalente en español al marcador adversativo "en cambio", hace que el A2 etiquete este fragmento con una relación de Contraste, mientras que el A1, al no contar con ningún tipo de marcador del discurso, lo ha etiquetado como una relación de Lista.

d) Suprimir oraciones de relativo:

> 25a. [Hemos estudiado retrospectivamente 23 infecciones protésicas de rodilla tratadas en nuestro hospital entre el año 1996 y el 2004 *de las cuales* hemos excluido 6 por diferentes motivos.]
>
> 25b. [1996. eta 2004. urteen bitartean gure ospitalean izandako 23 infekzio protesiko aztertu ditugu.] [*Horien artean*, 6 kasu *baztertu ditugu* hainbat arrazoirengatik.]

En este ejemplo, en euskera se ha evitado la estructura de relativo del fragmento en español y se ha traducido mediante una oración independiente con un marcador anafórico del discurso ("horien artean") para recuperar la fuerza de la estructura relativa del español.

Una vez observados todos los casos, concluimos que el uso de las técnicas de traducción detectadas se debe a que en euskera la oración en su orden neutro lleva la carga al final, ya que sigue la estructura SOV (Sujeto-Objeto-Verbo). Para facilitar la comprensión se adelanta la carga de la oración o se reduce su tamaño, utilizando así más oraciones para traducir el mismo contenido semántico. Precisamente por esta razón, para acortar las oraciones, se han utilizado diversas técnicas de traducción en euskera. El uso de estas técnicas afectará además a la estructura retórica de un modo definitivo, cambiando las relaciones entre las EDUs y por lo tanto cambiando el significado del texto. El uso de estas técnicas también aumenta el desacuerdo entre los anotadores sobre las relaciones porque, además de acortar las oraciones, también se cambia la semántica de las relaciones y, por lo tanto, la interpretación que se hace de ellas.

En la Tabla 8 se muestran los datos de la evaluación parcial cualitativa realizada en este trabajo.

	Números absolutos	**Porcentajes**
Total de relaciones	224	100%
Acuerdo sobre relaciones	157	71%
Desacuerdo sobre relaciones	65	29%
Desacuerdos por traducción	31	13,8%
Desacuerdos no justificados	34	15,2%

Tabla 8. Datos de la evaluación parcial cualitativa

Finalmente, en la Tabla 9 se ofrecen la precisión y la cobertura de la evaluación cuantitativa, y la precisión de la evaluación cualitativa. Se observa que la precisión en ambas evaluaciones es muy similar (un 73,42% en la cuantitativa y un 74.9% en la cualitativa).

	Cuantitativa		**Cualitativa**
	Total cobertura	**Total precisión**	**Total precisión**
Relaciones	98.6%	73.42 %	74.09 %

Tabla 9. Resultados finales de la evaluación cuantitativa y de la evaluación parcial cualitativa

Como se observa en la Tabla 9, la precisión de la evaluación cualitativa de la comparación de los 20 árboles retóricos es algo más optimista que la cuantitativa,

aunque no demasiado. Sin embargo esta situación no es constante, ya que en algunos árboles la diferencia entre las evaluaciones oscila aproximadamente entre el +/- 10%.

4. Conclusiones

A modo de conclusión general, creemos que este trabajo supone una nueva contribución en relación con la RST, ya que amplía el estado de la cuestión sobre la comparación de árboles retóricos en diferentes lenguas, como por ejemplo el de Marcu et al. (2000), en inglés y japonés. Se han mencionado además algunos problemas de la evaluación cuantitativa y se ha presentado una evaluación cualitativa. Asimismo, nuestro trabajo demuestra que, aunque existen diferencias en el análisis retórico realizado por dos anotadores sobre un mismo corpus (con textos paralelos en dos lenguas), especialmente debido a las técnicas de traducción empleadas, hay menos subjetividad de la que se podría esperar. Por tanto, consideramos que la utilización de la RST para múltiples finalidades, tanto teóricas como aplicadas, queda ampliamente justificada.

Otra de las conclusiones del trabajo es que las técnicas de traducción empleadas influyen en la interpretación de las relaciones retóricas de la RST. El traductor (en este caso el mismo médico que redacta el artículo de investigación, por tanto no puede decirse que se trate realmente de un traductor especializado) en ocasiones no emplea las mismas estructuras lingüísticas al traducir de una lengua a otra. Este hecho ha sido determinante para que los dos anotadores de nuestro estudio interpretasen de manera diferente un mismo fragmento redactado en dos lenguas diferentes.

Asimismo, la comparación de los árboles retóricos de textos paralelos nos ha permitido constatar dos situaciones: a) a la hora de traducir un resumen, no se tiene tanto en cuenta la estructura retórica como la estructura sintáctica y b) en los casos en que las estructuras retóricas son imposibles de traducir por analogía, las técnicas de traducción empleadas nos dan pistas sobre las tendencias de las lenguas a estructurar el discurso (lo cual es una cuestión a tener en cuenta en la traducción automática de las estructuras retóricas).

Como trabajo futuro nos plantemos indagar en las razones de las oscilaciones entre la evaluación cuantitativa y cualitativa y además añadir a este estudio una tercera lengua, el inglés, ya que, como ya hemos comentado, la Gaceta Médica de Bilbao incluye también los resúmenes de los autores en esta lengua. Consideramos de gran interés observar qué tipos de técnicas se han empleado para la traducción del discurso y qué diferencias existen entre ellas. Siendo el inglés y el español lingüísticamente más cercanos, deberían reducirse las técnicas de traducción aplicadas y, por lo tanto, esta variable se vería disminuida al comparar lenguas más cercanas. También en el futuro nos planteamos realizar un repositorio de marcadores discursivos de relaciones retóricas en español, euskera e inglés a partir del análisis del corpus de resúmenes médicos redactados en estas tres lenguas para seguir estudiando, como en Iruskieta et al. (en prensa), las correlaciones entre las relaciones retóricas y los marcadores del discurso.

Bibliografía

Bouayad-Agha, N. (2000). "Using an abstract rhetorical representation to generate a variety of pragmatically congruent texts". *Proceedings of the 38th Meeting of the Assotiation for Computational Linguistics*. *Student Workshop*. 16-22.

Burstein, J.; Marcu, D. (2003). “A machine learning approach for identification of thesis and conclusion statements in student essays”. *Computers and the Humanities* 37 (4). 455-467.

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

Carlson, L.; Marcu, D.; Okurowski, M. E. (2001). “Building a discourse-tagged corpus in the framework of rhetorical structure theory”. *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue*. 1-10.

da Cunha, I. (2008). *Hacia un modelo lingüístico de resumen automático de artículos médicos en español*. Barcelona: IULA. [CD-ROM] (Sèrie Tesis; 23)

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

Ghorbel, H.; Ballim, A; Coray, G. (2001). “ROSETTA: Rhetorical and Semantic Environment for Text Alignment”. En Rayson. P.; Wilson, A.; McEnery, A. M.; Hardie A.; Khoja, S. (eds.). *Proceedings of Corpus Linguistics 2001*. 224-233.

Haouam, K.; Marir, F. (2003). “SEMIR: Semantic indexing and retrieving web document using Rhetorical Structure Theory”. *Lecture Notes in Computer Science*. 596-604.

Iruskieta, M.; Diaz de Ilarraza, A.; Lersundi, M. (en prensa). “Correlaciones en euskera entre las relaciones retóricas y los marcadores del discurso”. *Actas del XXVII Congreso Internacional de AESLA: Modos y formas de la comunicación humana*. Ciudad Real: Universidad de Castilla-La Mancha.

Mann, W. C. (2005). *RST Web Site*. www.sfu.ca/rst [Consulta: 15/08/2009]

Mann, W. C.; Thompson, S. A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. ISI: Information Sciences Institute, Los Ángeles, CA, ISI/RS-87-190, 1-81.

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

Marcu, D. (1998). The rhetorical parsing, summarization, and generation of natural language texts. Toronto, University of Toronto. [Tesis doctoral]

Marcu, D.; Amorrortu, E.; Romera, M. (1999). "Experiments in constructing a corpus of discourse trees”. *Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging*. 48-57.

Marcu, D. (2000a). *The Theory and Practice of Discourse Parsing Summarization*. Massachusetts: Institute of Technology.

Marcu, D. (2000b). “The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach”. *Computational Linguistics* 26 (3). 395-448.

Marcu, D.; Carlson, L.; Watanabe, M. (2000). "The automatic translation of discourse structures”. *Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics*. 9-17.

O’Donnell, M. (2000). “RSTTOOL 2.4 – A Markup Tool for Rhetorical Structure Theory”. *Proceedings of the International Natural Language Generation Conference*. 253-256.

Pardo, T.A.S.; Nunes, M.G.V.; Rino, L.H.M. (2004). “DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese”. *Lecture Notes in Artificial Intelligence*. 224-234.

Pardo, T.A.S.; Nunes, M.G.V. (2008). “On the Development and Evaluation of a Brazilian Portuguese Discourse Parser”. *Journal of Theoretical and Applied Computing*. 15 (2). 43-64.

Stede, M. (2008). “Disambiguating Rhetorical Structure”. *Journal of Research in Language and Computation* 6. 311-332.

Sumita, K.; Ono, K.; Chino, T.; Ukita, T.; Amano, S. (1992). “A discourse structure analyzer for Japonese text”. *Proceedings of the International Conference on Fifth Generation Computer Systems*. 1133-1140.

Taboada, M.; Mann, W.C. (2005). “Applications of rhetorical structure theory”. *Discourse Studies*. 8 (4). 567-588.

Iria da Cunha Fanego pertenece al grupo de investigación IULATERM del Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra (UPF) de Barcelona. Actualmente se encuentra realizando una estancia postdoctoral en el grupo TALNE (Traitement Automatique du Langage Natural Écrit) del Laboratoire Informatique d'Avignon, financiada por el Ministerio de Ciencia e Innovación de España mediante el Programa Nacional de Movilidad de Recursos Humanos del Plan nacional de I-D+I 2008-2011.
Páginas web:
http://www.iula.upf.edu/
http://lia.univ-avignon.fr/

Mikel Iruskieta Quintian pertenece al grupo de investigación IXA de la Facultad de Informática de la Universidad del País Vasco (UPV-EHU). Actualmente es profesor en el Departamento de Didáctica de la Lengua y la Literatura en la Escuela Universitaria de Magisterio de Bilbao.
Página web:
http://ixa.si.ehu.es/Ixa