Vol. 8, n. 3, p. 181-202, set/dez 2010
 - doi: 10.4013/cld.2010.83.03

Mikel Iruskieta
mikel.iruskieta@gmail.com

Iria da Cunha
iria.dacunha@upf.edu

El potencial de las relaciones retóricas para la discriminación de textos especializados de diferentes dominios en euskera y español

Potential of rhetorical relations for differentiation among specialized texts from different domains in Basque and Spanish

RESUMÉN – En este trabajo presentamos un estudio realizado con el fin de averiguar si las relaciones retóricas y las marcas superficiales que las evidencian tienen potencial para distinguir entre textos especializados de diferentes ámbitos que comparten un nivel de especialización alto, en dos lenguas tan diferentes como el euskera y el español. Para el análisis, hemos partido de la *Rhetorical Structure Theory* (RST). Hemos conformado un corpus paralelo de textos especializados español-euskera que contiene dos subcorpus, que incluyen textos del ámbito médico y del ámbito terminológico. Hemos anotado los textos con las relaciones retóricas de la RST y hemos detectado los marcadores del discurso que las evidencian. Finalmente, hemos observado que ciertas relaciones retóricas y la cantidad de marcadores del discurso empleados permiten discriminar un subcorpus de otro, tanto en euskera como en español.

Palavras-chave: *Rhetorical Structure Theory*, relaciones retóricas, marcadores del discurso, anotación, texto especializado, estudio contrastivo, español, euskera

ABSTRACT – This study presents our research on the potential of using rhetorical relations and superficial marks evidencing them to discriminate among specialized texts of different domains but with a high specialization level, in two very different languages as Basque and Spanish. For our analysis, we employ of the Rhetorical Structure Theory (RST). We compiled a parallel corpus of Spanish-Basque specialized texts that contains two subcorpora of medical and terminological texts. We marked these texts with RST rhetorical relations and we detected the discourse markers that evidence them. Finally, we noted that certain types of rhetorical relations and the amount of used discourse markers allow us to differentiate among specialized texts of different domains in both Spanish and Basque.

Key words: Rhetorical Structure Theory, rhetorical relations, discourse markers, annotation, specialized text, contrastive study, Spanish, Basque.

Introducción

Hoy en día, el análisis del discurso es un ámbito muy estudiado. En concreto, el análisis de la estructura retórica de documentos ha suscitado gran interés. Una de las teorías más empleadas para el análisis de la estructura retórica es la *Rhetorical Structure Theory* (RST) de Mann y Thompson (1988), una teoría organizativa del texto que describe su estructura a partir de las relaciones que se establecen entre sus diferentes elementos discursivos[1] (núcleos y satélites).

La mayor parte de los trabajos sobre el análisis de la estructura retórica se han realizado sobre corpus formados por textos monolingües. No obstante, hay algunos autores que sí han llevado a cabo estudios en donde se comparan estructuras retóricas de lenguas diferentes, como chino-inglés (Cui, 1986; Kong, 1998; Ramsay, 2000, 2001), inglés-holandés (Abelen *et al.*, 1993), inglés-francés (Delin *et al.*, 1996; Salkie y Oates, 1999), portugués-francés-inglés (Scott *et al.*, 1998), inglés-japonés (Marcu *et al.*, 2000) o español-euskera (da Cunha e Iruskieta, 2010). La mayor parte de estos estudios se realizaron con el objetivo de analizar las diferencias existentes entre las estructuras retóricas de dos o más lenguas, empleando para ello corpus paralelos que contenían textos de una sola temática o de ámbito general.

[1] En este trabajo emplearemos los adjetivos "retórico" y "discursivo" como sinónimos.

También el discurso especializado es un tema muy estudiado desde hace años, desde diferentes puntos de vista y con diferentes objetivos. Por ejemplo, hay trabajos que han detectado rasgos específicos de géneros diferentes (véase Swales, 1981, 1990; van Dijk, 1980, 1989; Kaplan *et al.*, 1994, Ciapuscio, 1998) o diferencias entre textos especializados y textos de ámbito general (véase, L'Homme, 1992; Teufel y Moens, 2002; Cabré, 2007; Cabré *et al.*, 2010). Para discriminar textos especializados de diferentes ámbitos, lo más habitual es utilizar rasgos léxicos, en especial, la terminología (Cabré y Estopá, 2003; Ciapuscio, 2003). Para caracterizar un tipo de género, pueden utilizarse asimismo rasgos gramaticales (véase Graetz, 1985). Para detectar rasgos específicos de un género asociado a un ámbito especializado, también puede emplearse la estructura textual (véase Salager-Meyer, 1991). También existen trabajos que muestran ciertas características del discurso especializado (Rey, 1999) y de la comunicación científica (Jacobi, 1999), del discurso de divulgación científica en general (Rey y Tricas, 2006; Mortureux, 1988) e incluso del discurso de divulgación en el ámbito médico (Olivares, 2006), así como obras que reflejan la tipología del discurso científico (Loffler-Laurian, 1983). En el trabajo de Pérez (2001) se realiza un análisis retórico contrastivo de resúmenes lingüísticos y médicos en inglés y español. Sin embargo, no encontramos trabajos que exploten la estructura retórica de la RST para discriminar entre textos de dominios especializados diferentes pero de un mismo nivel de especialización, ni tampoco trabajos sobre este tema para lenguas diferentes al español, inglés o francés. Con respecto al nivel de especialización, consideramos, siguiendo a Cabré (1998), que existen tres: nivel alto (emisor experto y receptor experto), nivel medio (emisor experto y receptor aprendiz de la materia) y nivel bajo (emisor experto y receptor lego en la materia).

El objetivo de este trabajo es analizar el potencial que tienen las relaciones retóricas de la RST para discriminar textos especializados de diferentes dominios pero de un mismo nivel de especialización (alto) y de un mismo género (resumen de artículo de investigación), y observar si este potencial es el mismo en diferentes lenguas, en concreto en español y en euskera. Para ello, hemos conformado un corpus paralelo de resúmenes de artículos de investigación en estas dos lenguas, de dos ámbitos especializados muy diferentes entre sí: el ámbito médico y el ámbito terminológico. Hemos seleccionado la medicina y la terminología porque consideramos que son dos ámbitos muy diferentes (ciencias médicas vs. ciencias humanas) y presuponemos que los textos que escriben los especialistas de dichos ámbitos pueden tener unas características discursivas propias que puedan distinguirlos, incluso aunque se trate de textos del mismo género textual y del mismo nivel de especialización. Así, en este trabajo, hemos llevado a cabo el análisis de la estructura retórica de este corpus paralelo bilingüe de carácter ejemplar para intentar detectar la existencia de rasgos discursivos que puedan ayudar a diferenciar textos de diferentes ámbitos especializados.

Este trabajo lingüístico, de carácter descriptivo y analítico, puede servir asimismo como base de trabajos de lingüística computacional. Concretamente, puede ser útil en las áreas de extracción y recuperación de información, para optimizar sistemas de búsqueda de textos especializados, por ejemplo. Los resultados podrían aplicarse también al área de clasificación textual, ya que los rasgos discursivos específicos de cada dominio pueden tomarse como base de sistemas automáticos con el objetivo de determinar el ámbito al que pertenece un texto (o un conjunto de textos). Estos sistemas de clasificación son necesarios para constituir corpus especializados de manera automática, ya que el proceso de elaboración manual de este tipo de corpus es muy costoso. Los corpus especializados son imprescindibles para realizar diversas tareas, como la elaboración de diccionarios o léxicos especializados que contengan la terminología de un dominio, o la construcción de ontologías o taxonomías que estructuren el conocimiento de un dominio.

En el segundo apartado detallamos la metodología del estudio. En el tercer apartado exponemos el análisis realizado y los resultados obtenidos. En el cuarto apartado presentamos las conclusiones del trabajo y el trabajo futuro.

Metodología

La metodología que hemos empleado en este trabajo tiene varias fases, que detallamos en este apartado.

Teoría empleada

En primer lugar, seleccionamos la teoría del discurso más adecuada para la anotación de nuestro corpus. Como ya hemos adelantado en la introducción, empleamos la RST, por ser una teoría muy útil para describir un documento caracterizando su estructura mediante las relaciones que se establecen entre las diferentes unidades discursivas del mismo. Estas unidades pueden ser de dos tipos: "núcleo" (si es más relevante en relación con el objetivo del emisor del texto) o "satélite" (si ofrece una información retórica acerca de algún núcleo). La estructura retórica más habitual en la lengua es la de "núcleo-satélite". Relaciones de este tipo son Elaboración, Concesión, Antítesis, Condición, Reformulación, Propósito, etc. Sin embargo, también existen estructuras en las que existen diversos núcleos relevantes para el objetivo del emisor; son las estructuras "multinucleares". Relaciones de este tipo son, por ejemplo, Lista, Secuencia y Contraste, entre otras. Para una explicación detallada de la RST, remitimos a los artículos de Mann y Thompson (1988), y Mann y Taboada (2010).

Decisiones metodológicas en cuanto a la anotación

En segundo lugar, tomamos algunas decisiones metodológicas en cuanto a la anotación discursiva. Por un lado, determinamos la lista de relaciones retóricas empleadas: la anotación extendida de Mann y Taboada (2010). En el Anexo 1 se ofrece la lista de estas relaciones y se especifica si son relaciones núcleo-satélite (N-S) o multinucleares (N-N).

Por otro lado, con respecto a la segmentación de las Unidades Discursivas Mínimas (*Elemental Discourse Units*, EDUs), seguimos las pautas ofrecidas originalmente por la RST, aunque con algunos matices. Estos matices son los mismos que se recogen en los trabajos de da Cunha e Iruskieta (2010) y Tofilosky *et al.* (2009), y se refieren básicamente a tres aspectos: (i) no se consideran como EDUs los fragmentos que no contengan verbos finitos[2], (ii) que constituyan una cláusula de relativo o (iii) que sean cláusulas subordinadas de sujeto, de complemento directo o completivas. Así, segmentaríamos los siguientes fragmentos, que se corresponden con cada uno de estos tres aspectos, de la siguiente manera[3]:

(1a) [Los ejes de nuestro estudio, tal y como hemos indicado más arriba, son los siguientes: *lograr un discurso jurídico correcto y de calidad, tanto desde el punto de vista del derecho como del lingüístico, utilizando la traducción y la terminología en la configuración de dicho discurso.*]$_{EDU1}$

(2b) [Todos estos factores, además de provocar un aumento cuantitativo de la terminología especializada, han implicado una ampliación de la perspectiva del trabajo en terminología, *que si bien la ha enriquecido, al mismo tiempo ha puesto en cuestión algunos de sus conceptos básicos,* […].]$_{EDU1}$

(3a) [Es importante asegurarnos *de que los términos acuñados son creados sistemáticamente, no son ambiguos ni en su significado ni en su uso y sí consecuentes con otros términos relacionados del mismo campo.*]$_{EDU1}$

También decidimos emplear una convención que llamamos Same-Unit en los casos en que una EDU esté truncada por otra, es decir, en el caso de que una EDU contenga otra en su interior, siguiendo la línea de Carlson y Marcu (2001). Las Figuras 1 y 2 muestran un ejemplo de anotación de la relación Same-Unit en español y en euskera, respectivamente.

(4a) [En décadas precedentes se ha puesto de manifiesto,] [y así lo han atestiguado muchos investigadores de la terminología científica serbia,] [una tendencia a importar préstamos de unidades estructurales tanto léxicas como otras mayores del inglés a una serie de registros científicos específicos, en lugar de optar por la traducción, el calco, etc.

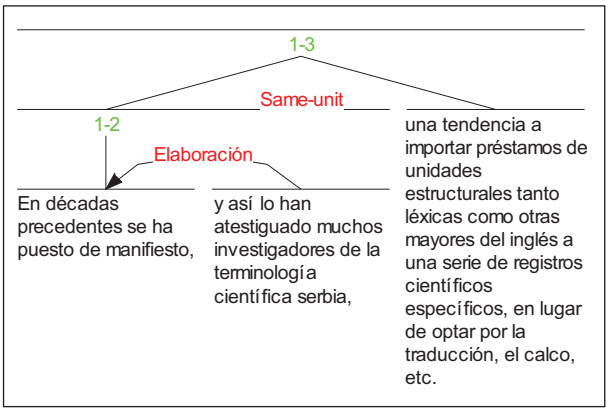

Figura 1. Fragmento de árbol retórico en español que incluye la relación Same-Unit.
Figure 1. Rhetorical tree in Spanish including the Same-Unit relation.

(4b) [Aurreko hamarkadetan, serbierako zientzia-arloko ikertzaile askok joera bat nabaritu dute eta] [horren berri eman dute:] [ingeleseko unitate lexikalen maileguak eta unitate-egitura luzeagoen maileguak hartzen dira zientzia-erregistro zehatz baterako, itzulpenak edo kalkoak egin ordez.]

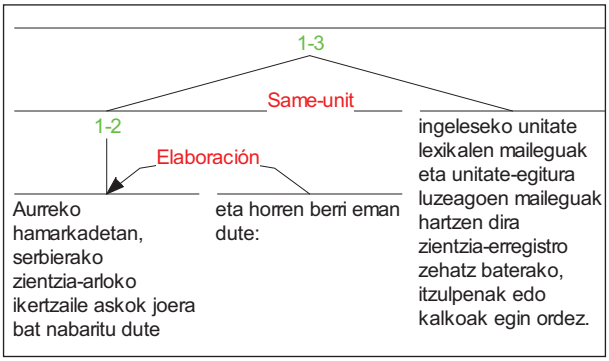

Figura 2. Fragmento de árbol retórico en euskera que incluye la relación Same-Unit.
Figure 2. Rhetorical tree in Basque including the Same-Unit relation.

[2] A excepción de los títulos, que sí pueden constituir una EDU sin incluir un verbo finito.
[3] Todos los ejemplos ofrecidos en este artículo provienen del corpus que hemos analizado en nuestro estudio. Las EDUs analizadas en los ejemplos están marcadas entre corchetes.

Corpus

En tercer lugar, conformamos un corpus paralelo formado por resúmenes de artículos de investigación en euskera y en español. En la actualidad, es difícil encontrar corpus paralelos español-euskera disponibles de gran tamaño para la comunidad científica y, por este motivo, tuvimos que diseñar y recopilar nuestro propio corpus de carácter ejemplar para su estudio retórico. En el País Vasco es habitual que las revistas de investigación o las actas de congresos incluyan los resúmenes de los textos publicados tanto en español como en euskera, así que este fue el recurso que empleamos. Conformamos un corpus que a su vez dividimos en dos subcorpus: un subcorpus formado por resúmenes de artículos médicos y un subcorpus formado por resúmenes de artículos sobre terminología. Los textos del corpus médico se extrajeron de la *Gaceta Médica de Bilbao*[4] y fueron escritos por especialistas entre los años 2000 y 2008. Los textos del corpus de terminología se extrajeron de las actas del Congreso Internacional de Terminología celebrado en Donostia y Gasteiz en 1997[5].

El objetivo comunicativo y el ámbito de uso de los textos de ambos subcorpus es el mismo: al tratarse de resúmenes tienen el propósito de informar en medios de comunicación especializados. Es decir, son textos informativos, ya que aportan de un modo resumido la información de una investigación en una revista especializada y en las actas de un congreso. Sin embargo, existen algunas diferencias en cuanto a la función de los textos incluidos en cada subcorpus: mientras que los resúmenes de los textos médicos tienen una función referencial, donde lo importante es presentar el objeto de estudio, los resúmenes de los textos terminológicos tienen, además, una función metalingüística. Asimismo, el objeto de estudio y el contexto es diferente en ambas áreas: el objeto de estudio de los resúmenes terminológicos está en relación con tareas de recolección, descripción y creación de términos de lenguas diversas en un congreso internacional, mientras que el objeto de estudio de los resúmenes médicos es la investigación sobre la salud/enfermedad del paciente en una revista local. Con respecto a la perspectiva desde la que se aborda cada área, es importante notar que la perspectiva de la terminología es explicativa-declarativa, ya que sus textos tienen un carácter definitorio e informativo (definen y dan a conocer una serie de hechos o datos), mientras que la medicina parte de una perspectiva explicativa-procedimental con una finalidad demostrativa (y los textos deben aportar pruebas, resultados de experimentos, datos estadísticos, etc. para defender y demostrar las afirmaciones realizadas).

El corpus en español incluye un subcorpus de 20 textos médicos y un subcorpus de 11 textos terminológicos. La diferencia en el número de textos se debe a que los resúmenes de artículos médicos suelen ser más breves que los resúmenes de artículos de terminología. Cada uno de estos dos subcorpus cuenta con aproximadamente 4000 palabras y 21000 caracteres. El corpus en euskera incluye los mismos textos en ambos subcorpus, ya que, de cara a nuestros fines de análisis retórico, necesitábamos que se tratase de un corpus paralelo bilingüe. Sin embargo, el número de palabras de los dos subcorpus en euskera es inferior al de los subcorpus en español (constan de unas 3000 palabras aproximadamente). Esta diferencia es lógica, ya que, a diferencia del español, el euskera es una lengua aglutinante y, por lo tanto, los artículos se insertan al final del sintagma y las preposiciones al final de la palabra que modifican. Por esto no se consideran como palabras independientes. Así, es normal que el corpus en español incluya un mayor número de palabras. Por esto, tal y como reflejan los datos de la Tabla 1, el número de caracteres de ambos corpus, español y euskera, es similar (en español: 21755 en el subcorpus médico y 20912 en el subcorpus terminológico/en euskera: 21530 en el subcorpus médico y 19774 en el subcorpus terminológico).

En el ejemplo 5a ofrecemos un fragmento del corpus español (en donde el segundo elemento constituye un satélite de Elaboración del primero) y en el ejemplo 5b mostramos su fragmento paralelo en euskera. En ambos fragmentos hemos subrayado preposiciones y artículos, que en español constituyen palabras independientes,

Tabla 1. Estadísticas del corpus.
Table 1. Corpus statistics.

	Textos ESP	Palabras ESP	Caracteres sin espacios ESP	Texto EUS	Palabras EUS	Caracteres sin espacios EUS
Subcorpus médico	20	4003	21755	20	3024	21530
Subcorpus terminológico	11	3724	20912	11	2563	19774

[4] http://www.gacetamedicabilbao.org/web/es/
[5] http://www.uzei.com/antcatalogo.asp?nombre=1687&hoja=0&sesion=14

mientras que en euskera, al ser una lengua aglutinante, se unen a otras palabras como sufijos.

(5a) [La informática jurídica documental, gira alrededor de las llamadas Bases de datos jurídicos.]$_{EDU1}$ [Las mismas presentan fundamentalmente tres tipos de documentos que responden a las tres fuentes principales del Derecho: jurisprudencia, legislación y doctrina de los autores.]$_{EDU2}$
(5b) [Agirietako informatika juridikoa datu-base juridikoen inguruan dago oinarrituta.]$_{EDU1}$ [Datu-base horietan hiru agiri-mota aurkitzen dira batez ere, zuzenbidearen hiru iturri nagusien arabera, hain zuzen: jurisprudentzia, legegintza eta autoreen doktrina.]$_{EDU2}$

En el Anexo 2 se ofrece una tabla con los datos de los textos del corpus: título, autor(es) y año de publicación.

Análisis discursivo

En cuarto lugar, realizamos el análisis discursivo del corpus mediante la anotación de las relaciones de la RST. El anotador 1 (A1) analizó los textos del corpus en español, mientras que el anotador 2 (A2) analizó los textos del corpus en euskera, de manera independiente y sin consultas entre ellos.

La anotación tuvo dos fases principales: (a) la segmentación de las EDUs y (b) el análisis retórico. Una vez finalizada la primera fase, se decidió homogeneizar las segmentaciones realizadas por ambos anotadores. Esta segmentación se llevó a cabo para minimizar el "ruido" que estas diferencias podrían provocar posteriormente de cara a la anotación discursiva, ya que, si ambos anotadores parten de segmentaciones diferentes, es muy complicado evaluar los análisis retóricos finales.

Una vez homogenizadas las segmentaciones del corpus en español y en euskera, los anotadores etiquetaron los textos asignando relaciones retóricas entre las EDUs y determinando cuáles de estas eran núcleos y satélites. Para realizar tanto la segmentación como la anotación discursiva, se empleó la herramienta RSTTool (O'Donnell, 2000).[6]

La Figura 3 muestra un árbol discursivo de un fragmento de un texto del subcorpus médico en español y en euskera, y la Figura 4 muestra un árbol discursivo de un fragmento de uno de los textos del subcorpus terminológico también en ambas lenguas.

Análisis cuantitativo

En quinto lugar, contabilizamos las relaciones discursivas de cada tipo y los marcadores discursivos detectados en el corpus en español y en euskera, tanto en el subcorpus médico como en el subcorpus terminológico. Tomamos el término "marcador discursivo" en un sentido amplio, es decir, considerando cualquier marca morfosintáctica que pueda ser indicadora de una relación discursiva, sin restringirnos a clasificaciones existentes, como serían las de Portolés (2001) o Montolío (2001). Ejemplos de marcadores serían: "porque", "por lo tanto", "sin embargo", "y", "de manera que", "con esto en mente", etc.

Finalmente, a partir de los resultados obtenidos, extraemos algunas conclusiones con respecto a las diferencias discursivas (tanto referentes a relaciones retóricas como a marcadores discursivos) detectadas en el corpus de medicina y el corpus de terminología, comparando los resultados obtenidos por ambas lenguas. Estas diferencias discursivas nos permitieron establecer algunas hipótesis sobre el potencial que tienen las relaciones retóricas para discriminar entre textos de ámbitos especializados diferentes y sobre los paralelismos existentes entre el español y el euskera.

Análisis y resultados

El análisis del corpus se realizó de la manera explicitada en el apartado anterior. Con respecto a la homogeneización de la segmentación, hemos detectado dos casos principales que la han motivado:

(i) *Diferencias de puntuación*. En ciertos casos, una puntuación diferente en un fragmento en español y euskera motiva una segmentación diferente por parte de los dos anotadores. El ejemplo 6a muestra un fragmento del corpus en español, con una cláusula de relativo ("la cual [...] política."), que el A1 anotó como una única EDU, ya que uno de los criterios de partida para nuestra anotación especifica que no deben segmentarse las cláusulas de relativo. El ejemplo 6b muestra el fragmento paralelo en euskera, donde esa cláusula de relativo se ha transformado en una oración diferente mediante la utilización del punto y coma, el marcador *era berean* ("así mismo") y el verbo finito *dago* ("está"). Por este motivo, y con la finalidad de homogeneizar, el A2 segmentó este fragmento en dos EDUs.

(6a) [El procesamiento cognitivo está, por lo tanto, inexorablemente unido a la gestión terminológica, la cual, por su parte, está unida a la planificación lingüística y a la política.]$_{EDU1}$
(6b) [Beraz, ezagutza-prozesaketak eta terminologia-kudeaketak daukaten lotura askaezina da;]$_{EDU1}$ [terminologia-kudeaketa, era berean, hizkuntz plangintzarekin eta politikarekin dago lotuta.]$_{EDU2}$

[6] http://www.wagsoft.com/RSTTool/

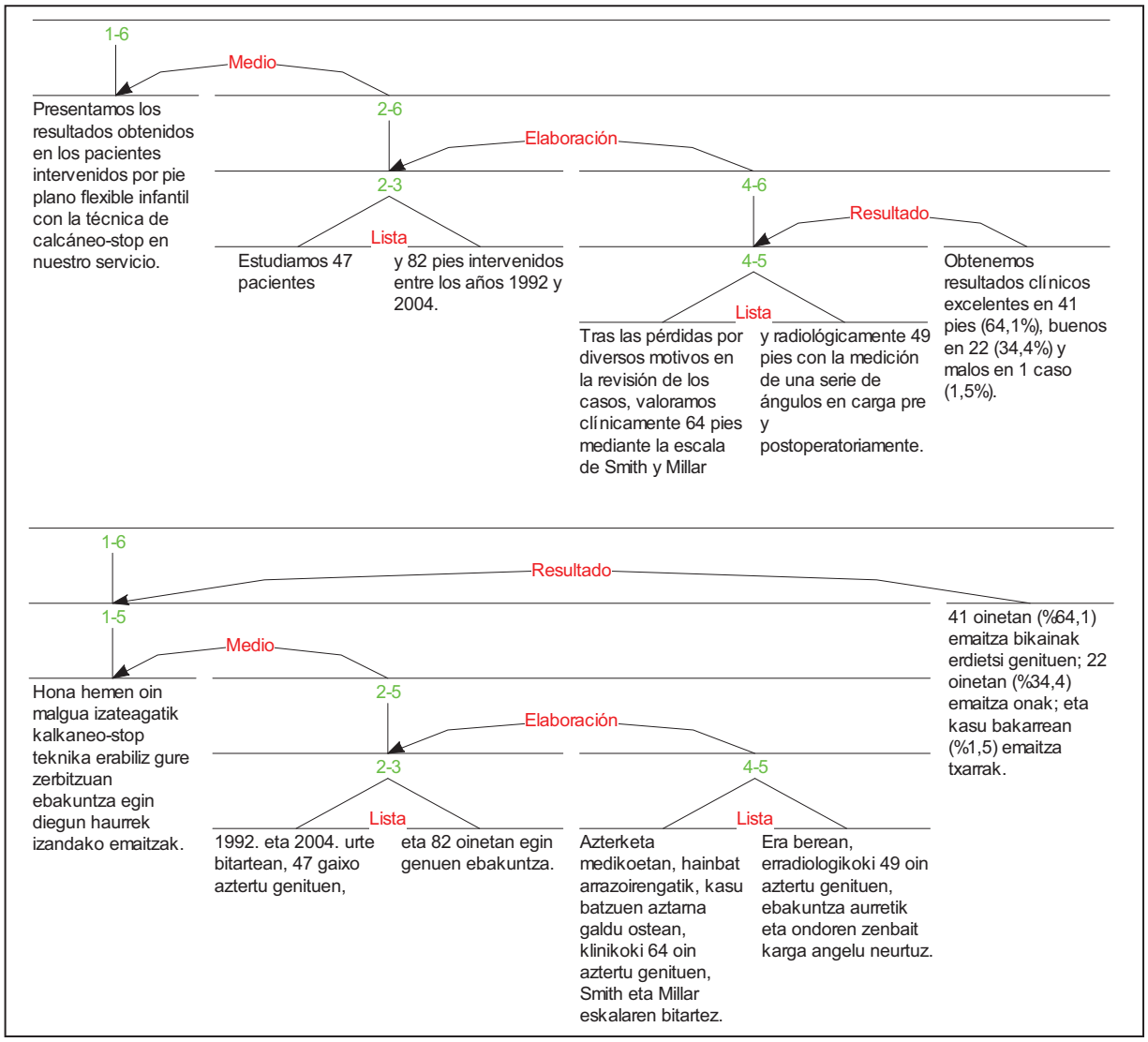

Figura 3. Árbol discursivo de un fragmento del corpus médico en español y en euskera.
Figure 3. Rhetorical tree of a medical corpus passage in Spanish and Basque.

(6c) TRADUCCIÓN: [Por lo tanto la unión entre el procesamiento cognitivo y la gestión terminológica es inexorable;]$_{EDU1}$ [la gestión terminológica, así mismo, está unida a la planificación lingüística y a la política.]$_{EDU2}$

En estos casos, homogeneizamos la segmentación en favor de la lengua en la que se haya segmentado más de una EDU, en este caso el euskera. Por tanto, la segmentación del fragmento 6a, se transformaría en la segmentación mostrada en el ejemplo 6d:

(6d) [El procesamiento cognitivo está, por lo tanto, inexorablemente unido a la gestión terminológica,]$_{EDU1}$ [la cual, por su parte, está unida a la planificación lingüística y a la política.]$_{EDU2}$

(ii) *Diferencias sintácticas*. En ocasiones, existen fragmentos paralelos español-euskera que cuentan con una estructura sintáctica diferente. Por este motivo, ambos anotadores pueden realizar segmentaciones diferentes. El ejemplo 7a muestra un fragmento del corpus en español con una estructura completiva ("asegurarse de" + complemento) que el A1 anotó como una única EDU, ya que uno de los criterios de partida para nuestra anotación especifica que no deben segmentarse este tipo de elementos sintácticos. El ejemplo 7b muestra el fragmento paralelo en euskera, que no incluye una cláusula completiva, sino una coordinación de tres oraciones con verbos finitos (*sortu behar dira*, "han de ser creados"; *ezin dira izan*, "no pueden ser"; *izan behar dira*, "deben ser"), que el A2 segmentó en tres EDUs:

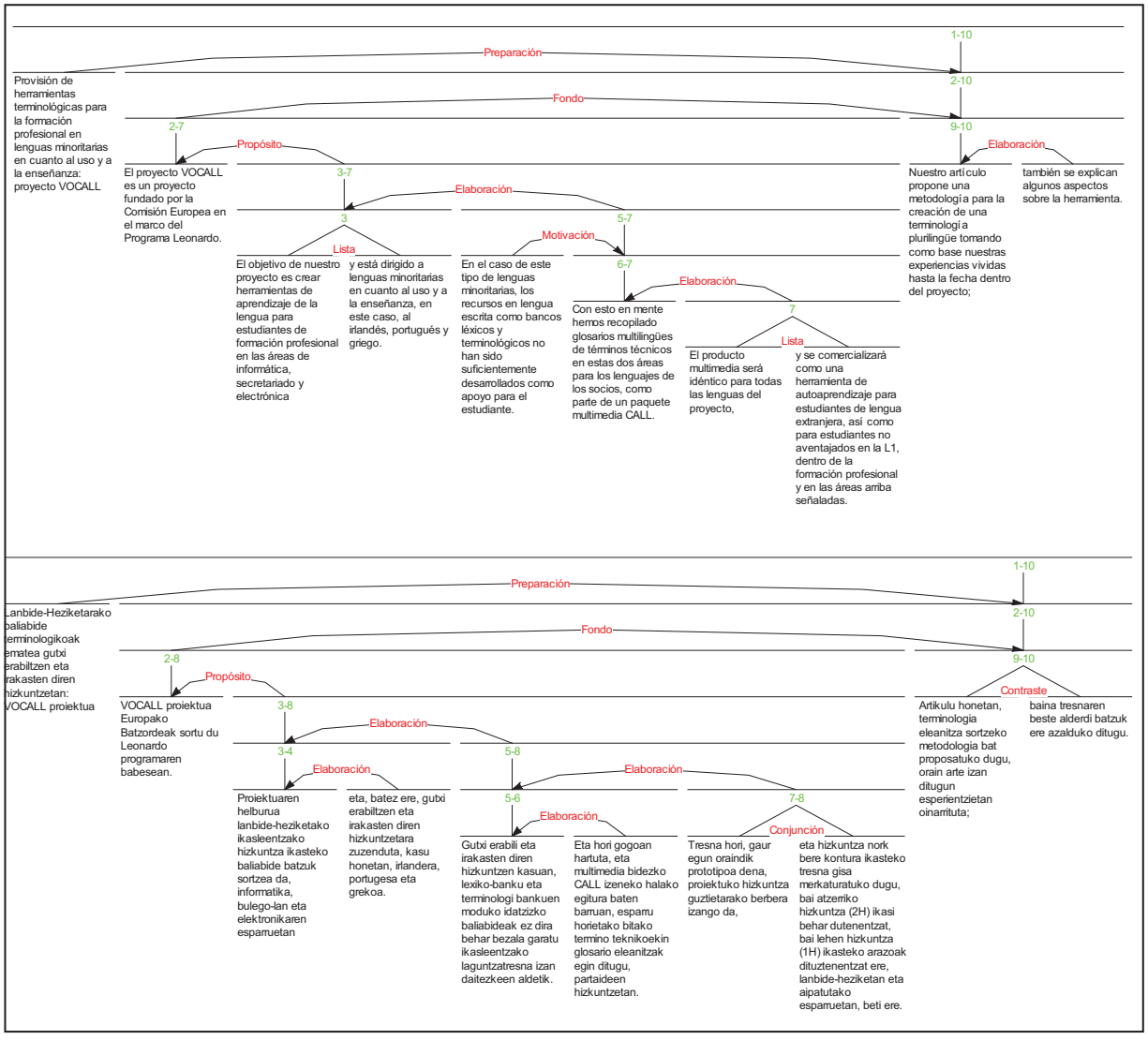

Figura 4. Árbol discursivo de un fragmento del corpus terminológico en español y en euskera.
Figure 4. Rhetorical tree of a terminological corpus passage in Spanish and Basque.

(7a) [Es importante asegurarnos de que los términos acuñados son creados sistemáticamente, no son ambiguos ni en su significado ni en su uso y sí consecuentes con otros términos relacionados del mismo campo.]$_{EDU1}$
(7b) [Asmatzen diren terminoak sistematikotasunez sortu behar dira,]$_{EDU1}$ [esanahiari eta erabilerari dagokionean ezin dira anbiguoak izan,]$_{EDU2}$ [arlo bereko pareko beste termino batzuekin koherenteak izan behar dute.]$_{EDU3}$]
(7c) TRADUCCIÓN: [Los términos acuñados han de ser creados sistemáticamente,]$_{EDU1}$ [no pueden ser ambiguos ni en su significado ni en su uso,]$_{EDU2}$ [deben ser consecuentes con otros términos relacionados del mismo campo.]$_{EDU3}$

En estos casos, de nuevo homogeneizamos la segmentación en favor de la lengua en la que se haya segmentado más de una EDU, en este caso el euskera. Así, la segmentación del fragmento 7a, se transformó en la segmentación mostrada en el ejemplo 7d:

(7d) [Es importante asegurarnos de que los términos acuñados son creados sistemáticamente,]$_{EDU1}$ [no son ambiguos ni en su significado ni en su uso]$_{EDU2}$ [y sí consecuentes con otros términos relacionados del mismo campo.]$_{EDU3}$

Una vez realizado el análisis discursivo, detectamos ciertas diferencias representativas entre los textos de ambos subcorpus. Por un lado, encontramos diferencias

relacionadas con las distintas modalidades de los textos: explicativa-procedimental (resúmenes médicos) y explicativa-declarativa (resúmenes terminológicos). La diferencia más relevante tiene que ver con la cantidad de relaciones discursivas de cada subcorpus: 192 en el corpus terminológico y 224 en el corpus médico, tanto en español como en euskera. Teniendo en cuenta que ambos subcorpus tienen un tamaño similar, esta diferencia de 32 relaciones significa que las unidades discursivas del subcorpus médico incluyen menos unidades léxicas, es decir, que para expresar una proposición emplean menos palabras. En la misma línea, hemos observado que en los resúmenes médicos se emplean menos marcas superficiales para señalar relaciones entre unidades discursivas, es decir, que las relaciones de coherencia entre las unidades retóricas no están señaladas de una manera tan marcada como lo están en los resúmenes terminológicos, como veremos más adelante.

También detectamos, tanto en español como en euskera, un mayor número de relaciones de Medio en el subcorpus médico que el subcorpus terminológico (en español: 19 en el subcorpus médico vs. 6 en el subcorpus terminológico / en euskera: 16 en el subcorpus médico vs. 8 en el subcorpus terminológico), aunque en este caso la diferencia es más significativa en el corpus en español. Estos datos suponen, en el corpus en español, el uso de un 8,48% de relaciones de este tipo en el subcorpus médico y un 3,12% en el subcorpus terminológico, mientras que en el corpus en euskera suponen un 7,14% y un 4,16%, respectivamente. Esta diferencia es debida a que la perspectiva del objeto en los textos médicos es explicativa-procedimental, mientras que en los textos terminológicos es más explicativa-declarativa. De ahí que en los resúmenes médicos se reflejen las herramientas y métodos utilizados en los experimentos en mayor medida que en los resúmenes sobre terminología, como puede apreciarse en los ejemplos 10a y 10b (en español y euskera, respectivamente), extraídos del subcorpus médico:

> (10a) [Este estudio tiene por objeto conocer datos relevantes en el proceso de atención del ictus en los hospitales públicos de Álava.]$_{\text{NÚCLEO}}$ [Se han recogido las demoras en la asistencia debidas al propio paciente o a su traslado a urgencias (tiempo entre el inicio de los síntomas y la entrada en urgencias) y también las debidas al sistema de atención (tiempo hasta el inicio de toma de constantes y tiempo hasta la realización de TAC). Otros datos recogidos se refieren a la realización de pruebas complementarias, tipo de hospitalización, calidad del informe de alta y valoración de secuelas.]$_{\text{SATÉLITE\_MEDIO}}$
>
> (10b) [Azterlan honek helbirutzat du Arabako ospitale publikoetan iktusa hartseko dauden prozesuan azpimarragarriak diren datuak ezagutzea.]$_{\text{NÚCLEO}}$ [Insistentzian egondako atzerapenak bildu dira, gaixo berari edo larrialdietara eramana izateari (sintomak hasi zirenetik eta larrialdietan sartu arteko denbora) zor zaizkionak, eta gainera, arreta sistemari dagozkionak (konstanteak hartzen hasi eta TACa egin zaio arteko denbora). Bildutako beste datu batzuk egindako proba osagarriei, ospitaleratze motari, alta txostenaren kalitateari eta ondorio txarren balorazioari dagozkie.]$_{\text{SATÉLITE\_MEDIO}}$

Por otro lado, detectamos diferencias importantes entre el subcorpus médico y el subcorpus terminológico en relación con las relaciones discursivas utilizadas, en ambas lenguas. En primer lugar, en el subcorpus médico se emplea un mayor número de relaciones de Resultado que en el subcorpus terminológico (en español: 17 en el subcorpus médico vs. 7 en el subcorpus terminológico / en euskera: 14 en el subcorpus médico vs. 4 en el subcorpus terminológico). Esto quiere decir que el corpus en español incluye un 7,58% de relaciones de Resultado en el subcorpus médico y un 3,64% en el subcorpus terminológico, mientras que el corpus en euskera contiene un 6,25% y un 2,08%, respectivamente. Estos datos confirman la línea argumental de la perspectiva del objeto, ya que en los textos explicativo-procedimentales se encuentran más relaciones de Resultado, ya que deben mostrar los resultados de una forma más tangible que los textos sobre estudios de terminología. El ejemplo 8a muestra un fragmento en el que se ha marcado una relación de Resultado en el subcorpus médico en español y el ejemplo 8b muestra el fragmento paralelo en euskera.

> (8a) [Realizamos asimismo un análisis comparativo entre el tamaño y forma de la lesión en el estudio ecográfico preoperatorio y el real de la intervención quirúrgica.]$_{\text{NÚCLEO}}$ [Pudimos aproximar de forma correcta el tamaño de la lesión con una tolerancia de error de 5 mm en el 89% de los casos, mientras que la información ecográfica properatoria de la forma de la lesión no se correspondía con la real intraoperatoria en más del 50% de los casos.]$_{\text{SATÉLITE\_RESULTADO}}$
>
> (8b) [Era berean, lesioaren neurriaren eta formaren arteko azterketa konparatiboa egin genuen ebakuntza aurreko azterlan ekografikoan eta ebakuntza kirurgiko berean.]$_{\text{NÚCLEO}}$ [Lesioaren neurria zehatz ezarri ahal izan genuen, 5 mm-ko errore tolerantziarekin kasuen %89an; lesioaren formako ebakuntza aurreko informazio ekografikoa ez da bat etortzen ebakuntza barneko kasuen %50ean.]$_{\text{SATÉLITE\_RESULTADO}}$

En segundo lugar, y en la misma línea, en el subcorpus médico se encuentran más relaciones de

Interpretación que en el subcorpus terminológico, de nuevo en ambas lenguas (en español: 13 en el subcorpus médico vs. 2 en el subcorpus terminológico / en euskera: 14 en el subcorpus médico vs. 1 en el subcorpus terminológico). Esto supone, en el corpus en español, la utilización de un 5,80% de relaciones de Resultado en el subcorpus médico y un 1,04% en el subcorpus terminológico; por su parte, en el corpus en euskera, supone un 6,25% y un 0,52%, respectivamente. Si en el subcorpus médico existen más relaciones de Resultado, es lógico pensar que habrá más relaciones de Interpretación sobre los resultados obtenidos. El ejemplo 9a muestra un fragmento en español en el que puede observarse una relación de Interpretación en el subcorpus médico; el ejemplo 9b muestra el fragmento paralelo en euskera.

(9a) [La aproximación al fenotipo "basal" definida según nuestros parámetros se correlacionó de manera altamente significativa con la expresión de p53 mutado ($p = 0.0001$), grado nuclear 3 ($p < 0.0001$) y un porcentaje de expresión de Ki67 igual o superior al 60% ($p < 0.0001$), y de manera apenas significativa con un grado histológico 3 ($p = 0.045$). No existió ningún grado de correlación con la invasión ganglionar ($p = 0.51$).] $_{\text{NÚCLEO}}$ [A falta de la determinación de citoqueratinas basales, la utilización de parámetros habitualmente presentes de forma rutinaria en los informes anatomopatológicos nos permite identificar un subgrupo de tumores "basal-like" ya en estadios muy precoces de la enfermedad, que se caracterizan por una elevada agresividad biológica.] $_{\text{SATÉLITE\_INTEPRETACIÓN}}$

(9b) [Gure parametroen arabera definitutako fenotipo "basala" korrelazioan jarri zen nabarmenki p53 mutatuaren agerpenarekin ($p = 0.0001$) 3. gradu nuklearra ($p < 0.0001$), eta %60 edo hortik gorako Ki67 agerpen ehunekoarekin; bestalde, korrelazioa oso txikia izan zen 3. gradu histologikoarekin ($p = 0.045$). Ganglioen inbasioarekin ($p = 0.51$) ez zen korrelazio gradurik hauteman.] $_{\text{NÚCLEO}}$ [Kitokeratina basalak zehaztu gabe daudenez, txosten anatomopatologikoetan erabili ohi diren parametroen bidez "basal-like" tumoreen azpitaldea hauteman dezakegu, gaitzaren egoera oso goiztiarrean. Azpitalde hori biologikoki oso erasokorra izan ohi da.] $_{\text{SATÉLITE\_INTEPRETACIÓN}}$

En tercer lugar, tanto en español como en euskera, detectamos un mayor número de relaciones de Contraste en el subcorpus médico que el subcorpus terminológico (en español: 12 en el subcorpus médico vs. 0 en el subcorpus terminológico/en euskera: 10 en el subcorpus médico vs. 2 en el subcorpus terminológico). Estos datos suponen, en el corpus en español, el uso de un 5,35% de relaciones de este tipo en el subcorpus médico y un 0,00% en el subcorpus terminológico, mientras que en el corpus en euskera suponen un 4,46% y un 1,04%, respectivamente. En los textos médicos, esta relación de Contraste se usa para demostrar la validez del estudio frente a otras investigaciones, y contraponer datos estadísticos o reacciones de pacientes, lo cual no ocurre en los textos terminológicos:

(11a) [El análisis estadístico reveló que los factores significativamente asociados a la invasión ganglionar fueron un grado histológico 3 ($p = 0.0066$) y un índice de Ki67 superior al 10% ($p = 0.0036$).] $_{\text{NÚCLEO\_CONTRASTE}}$ [Por el contrario, predijeron de manera significativa la ausencia de invasión ganglionar un tamaño de 5 mm o menor (pT1a) ($p = 0.022$), la variante histológica tubular pura ($p = 0.0026$), un grado nuclear 1 ($p = 0.0018$) y un grado histológico 1 ($p = 0.0022$).] $_{\text{NÚCLEO\_CONTRASTE}}$

(11b) [Azterketa estatistikoek erakutsi zuten gongoilaren inbasioari nabarmen lotutako faktoreak izan zirela; alde batetik, 3. maila histologikoa ($p = 0.0066$); eta bestetik, %10etik gorako Ki67 indizea ($p = 0.0036$).] $_{\text{NÚCLEO\_CONTRASTE}}$ Bestalde, gongoiletako inbasiorik ez zegoela nahiko ongi aurreikusi zen honako kasu hauetan: 5 mm edo txikiagoetan (pT1a) ($p = 0.022$), histologia barietate tubular puruan ($p = 0.0026$), 1. maila nuklearrean ($p = 0.0018$), eta 1. maila histologikoan ($p = 0.0022$).] $_{\text{NÚCLEO\_CONTRASTE}}$

En cuarto lugar, tanto en español como en euskera, detectamos un mayor número de relaciones de Concesión en el subcorpus terminológico que el subcorpus médico (en español: 6 en el subcorpus terminológico vs. 1 en el subcorpus médico / en euskera: 8 en el subcorpus terminológico vs. 2 en el subcorpus médico). En el corpus en español, hay un 3,12% de relaciones de este tipo en el subcorpus terminológico y un 0,44% en el subcorpus médico, mientras que en el corpus en euskera suponen un 4,16% y un 0,89%, respectivamente. Esta diferencia se debe no tanto al ámbito de uso, que es el mismo (medios de comunicación), sino al distinto objeto de estudio y al contexto. Los textos de terminología se presentan en un congreso internacional donde el objeto de estudio son distintas lenguas. Este contexto internacional y la diversificación del objeto de estudio (los lectores expertos, aunque no desconozcan las metodologías de trabajo, sí desconocen el objeto de estudio o lengua analizada) provocan que los terminólogos deban negociar más el significado. En cambio, el contexto de los textos del subcorpus médico es más local y el objeto no es tan disperso, sino que suele ser conocido por los lectores especializados:

(12a) [Es importante asegurarnos de que los términos acuñados son creados sistemáticamente, no son ambiguos ni en su significado ni en su uso y sí consecuentes con otros términos relacionados del mismo campo.]$_{\text{NÚCLEO}}$ [De cualquier forma, en el caso de cualquiera de las lenguas más importantes en que la uniformidad no sea una norma (por ejemplo: el chino que se habla en China continental, en Taiwan y en Hong Kong), la unificación de los términos empleados en una disciplina depende de la existencia de un banco terminológico dentro de una organización donde el aportar una guía para el uso, recopilación y mantenimiento de nuevos términos debería ser uno de sus deberes cotidianos.]$_{\text{SATÉLITE\_CONCESIÓN}}$

(12b) [Asmatzen diren terminoak sistematikotasunez sortu behar dira, esanahiari eta erabilerari dagokionean ezin dira anbiguoak izan, arlo bereko pareko beste termino batzuekin koherenteak izan behar dute.]$_{\text{NÚCLEO}}$ [Hala ere, hizkuntza handi batek ez badauka berdintasunik (adibidez, kontinenteko Txinan, Taiwanen eta Hong Kongen erabiltzen den txinerak), diziplina batean erabiltzen diren terminoak termino-banku baten bidez bateratu behar dira. Termino-banku hori termino berriak erabiltzeko, biltzeko eta gordetzeko irizpideak ematen dituen erakunde batean egon beharko luke.]$_{\text{SATÉLITE\_CONCESIÓN}}$

Detectamos, asimismo, como es de esperar en textos explicativos, que la relación más empleada en ambos corpus, español y euskera, como era de esperar ya que se trata de la relación más general, es la de Elaboración (en español: 56 en el subcorpus médico vs. 70 en el subcorpus terminológico/en euskera: 71 en el subcorpus médico vs. 59 en el subcorpus terminológico). Esto quiere decir que la relación de Elaboración supone, en el corpus español, un 25,00% de las relaciones empleadas en el subcorpus médico y un 36,45% en el subcorpus terminológico y, en euskera el subcorpus médico, un 31,69% y un 30,72% en el subcorpus terminológico, respectivamente. El ejemplo 13a muestra un fragmento en el que se ha anotado una relación de Elaboración en el subcorpus médico; el ejemplo 13b muestra el fragmento paralelo en euskera. Los ejemplos 14b y 14b, en español y euskera, incluyen también una relación de Elaboración, pero en este caso han sido extraídos del corpus terminológico.

(13a) [Se trata de tumores que no expresan receptores hormonales ni el oncogén c-erb-B2, y sí en cambio citoqueratinas propias de las células del estrato basal epitelial.]$_{\text{NÚCLEO}}$ [Dicho fenotipo tumoral, en consecuencia, se denomina "basal".]$_{\text{SATÉLITE\_ELABORACIÓN}}$

(13b) [Tumore horiek ez dituzte hormona hartzaileak eta c-erb-B2 onkogenea adierazten; eta bai, ordea, epitelio basaleko geruzaren zelulei dagozkien kitokeratinak.]$_{\text{NÚCLEO}}$ [Horren ondorioz, tumore fenotipo horri "basala" esaten zaio.]$_{\text{SATÉLITE\_ELABORACIÓN}}$

(14a) [En nuestro país, la única base de datos pública es el SISTEMA ARGENTINO DE INFORMÁTICA JURÍDICA, dependiente del Ministerio de Justicia de la Nación.]$_{\text{NÚCLEO}}$ [Esta base cuenta actualmente con más de 510.000 documentos.]$_{\text{SATÉLITE\_ELABORACIÓN}}$

(14b) [Gure herrialdean, datubase publiko bakarra INFORMATIKA JURIDIKOKO SISTEMA ARGENTINARRA da, Estatuko Justizi Ministerioaren mende dagoena.]$_{\text{NÚCLEO}}$ [Gaur egun, base horrek 510.000 agiri baino gehiago ditu.]$_{\text{SATÉLITE\_ELABORACIÓN}}$

En la Tabla 2 se muestran los datos cuantitativos del análisis discursivo realizado sobre nuestro corpus (corpus terminológico vs. corpus médico / ESP: español vs. EUS: euskera), nombre de la relación, tipo (N-N: relación Multinuclear; N-S: relación Núcleo-Satélite) y número de relaciones discursivas detectadas.

Con respecto a la cantidad de marcadores utilizados, existen asimismo algunas diferencias relevantes. En el corpus español, se encuentran 66 marcadores en el subcorpus terminológico y 48 en el subcorpus médico. En el corpus en euskera, se detectan 79 marcadores en el subcorpus terminológico y 70 en el subcorpus médico. Esta diferencia es significativa porque, si tenemos en cuenta que en el subcorpus médico en español y en euskera se emplea un 14,28% más de relaciones discursivas que en el subcorpus terminológico, esto quiere decir que el número de relaciones señaladas por marcadores discursivos en el subcorpus terminológico (en español: 34,37%/en euskera: 41,14%) es mucho más elevado que en el subcorpus médico (en español: 21,42%/en euskera: 31,25%). Debido al contexto internacional y a la dispersión del objeto analizado en los textos sobre terminología, además de negociar más el significado, es necesario señalar de manera más evidente las relaciones de coherencia.

En los Anexos 3 y 4 se ofrecen los datos cuantitativos del análisis de los marcadores en el corpus en español y en el corpus en euskera, respectivamente. Se indican el nombre de la relación, el tipo de relación, los marcadores detectados (con el número de ocurrencias de cada uno entre paréntesis) y el número total (T) de marcadores en ambos subcorpus (terminológico y médico). En el Anexo 5 se ofrece la traducción al español de los marcadores detectados en euskera.

Conclusiones

Este estudio constituye un primer análisis para detectar el potencial que tienen las relaciones discursivas de la RST para distinguir textos especializados que pertenecen a distintos ámbitos (en este estudio, médico vs. terminológico) pero que comparten un mismo nivel de especialización (concretamente, en nuestro trabajo, nivel alto) y un mismo género (en nuestro caso, el resumen de

Tabla 2. Análisis cuantitativo de las relaciones discursivas detectadas en el corpus.
Table 2. Quantitative analysis of the rhetorical relations detected into the corpus.

		Corpus terminológico		Corpus médico	
Relación	**Tipo**	**ESP**	**EUS**	**ESP**	**EUS**
Elaboración	N-S	70	59	56	71
Lista	N-N	31	15	59	46
Preparación	N-S	17	17	21	23
Fondo	N-S	13	14	10	11
Resultado	N-S	7	4	17	14
Justificación	N-S	6	1	0	5
Concesión	N-S	6	8	1	2
Medio	N-S	6	8	19	16
Propósito	N-S	6	6	2	4
Secuencia	N-N	4	12	2	0
Antítesis	N-S	4	1	2	1
Causa	N-S	4	13	4	1
Condición	N-S	3	5	2	2
Motivación	N-S	3	3	1	0
Interpretación	N-S	2	1	13	14
Circunstancia	N-S	2	2	2	1
Evaluación	N-S	2	0	0	0
Reformulación	N-S	1	3	0	0
Resumen	N-S	1	0	0	0
Solución	N-S	1	3	0	0
Disyunción	N-N	0	0	0	1
Capacitación	N-S	0	0	0	0
Conjunción	N-N	0	12	0	0
Evidencia	N-S	0	3	1	2
Alternativa	N-S	0	0	0	0
Unión	N-N	3	0	0	0
Contraste	N-N	0	2	12	10
Condición inversa	N-S	0	0	0	0
No-condicional	N-S	0	0	0	0
Total relaciones		**192**	**192**	**224**	**224**
Total N-N		**38**	**41**	**73**	**57**
Total N-S		**154**	**151**	**151**	**167**

artículo de investigación). Asimismo, hemos analizado si las diferencias observadas son similares en textos de lenguas diferentes; en concreto, hemos observado que los resultados del análisis son muy similares en el corpus paralelo español-euskera que hemos empleado. Por tanto, podría afirmarse que, en este estudio exploratorio, existe cierto potencial discriminatorio de las relaciones discursivas en ambas lenguas.

Los rasgos discursivos que según nuestro trabajo podrían ayudar a discriminar textos especializados de ámbitos diferentes tienen que ver con ciertas relaciones discursivas, como Resultado, Medio, Interpretación y Contraste, que aparecen en mayor medida en el corpus médico que en el corpus terminológico, tanto en español como en euskera. En cambio, en el subcorpus terminológico es la relación de Concesión la que podría ayudar a discriminar ambos tipos de textos. También hemos detectado que el número de marcadores en los textos sobre terminología es mucho más elevado que el de los textos médicos, también en ambas lenguas.

Para corroborar estas conclusiones será necesario realizar un estudio sobre un corpus textual más amplio,

es decir, que incluya un mayor número de textos, que nos permita obtener resultados estadísticamente significativos. Asimismo, nos gustaría realizar más experimentos empleando corpus de otros ámbitos especializados, como la economía, la informática, el medioambiente o el derecho. En esta misma línea, también creemos interesante realizar un estudio para analizar si es posible diferenciar textos especializados de textos de ámbito general teniendo en cuenta la aparición de ciertos rasgos discursivos.

Asimismo, consideramos que la metodología y los resultados de este trabajo pueden emplearse en investigaciones relacionadas con la extracción y recuperación de información de textos especializados, así como con la clasificación textual orientada a la constitución automática de corpus textuales especializados.

Referencias

ABELEN, E.; REDEKER G; THOMPSON. S.A. 1993. The rhetorical structure of US-American and Dutch fund-raising letters. *Text*, **13**(3): 323-350.

CABRÉ, M.T. 1998. Variació pel tema. El discurs especialitzat o la variació funcional determinada per la temática: noves perspectives. *Caplletra, Revista Internacional de Filología*, **25**:137-194.

CABRÉ, M.T. 2007. Constituir un corpus de textos de especialidad: condiciones y posibilidades. *In:* M. BALLARD; C. PINEIRA-TRESMONTANT (eds.), *Les corpus en linguistique et en traductologie*. Arras, Artois Presses Université, p. 89-106.

CABRÉ, M.T.; ESTOPÀ, R. 2003. On the units of specialised meaning uses in professional communication. *Terminology Science and Research*, **1**:217-237.

CABRÉ, M.T.; BACH, C.; DA CUNHA, I.; MORALES, A.; VIVALDI, J. 2010. Comparación de algunas características lingüísticas del discurso especializado frente al discurso general: el caso del discurso económico. *In: CONGRESO INTERNACIONAL DE AESLA: MODOS Y FORMAS DE LA COMUNICACIÓN HUMANA, XXVII,* Ciudad Real. *Anais...* Ciudad Real. Universidad de Castilla-La Mancha, p. 453-460.

CARLSON. L.; MARCU, D. 2001. *Discourse Tagging Reference Manual. ISI Technical Report ISITR-545.* Los Angeles, University of Southern California, 87 p.

CIAPUSCIO, G. 1998. Los resúmenes de la revista *Medicina*: Un enfoque diacrónico-contrastivo. *Revista Signo y Seña*, **10**:217-243.

CIAPUSCIO, G. 2003. *Textos especializados y terminología*. Barcelona, IULA, 149 p.

CUI, S. 1986. *A comparison of English and Chinese expository rhetorical structures*. Unpublished Master's thesis, UCLA.

DA CUNHA, I.; IRUSKIETA, M. 2010. Comparing rhetorical structures of different languages: The influence of translation strategies. *Discourse Studies*, **12**(5):563-598. http://dx.doi.org/10.1177/1461445610371054

DELIN, J.; HARTLEY, A.; SCOTT, D. 1996. Towards a contrastive pragmatics: Syntactic choice in English and French instructions. *Language Sciences*, **18**(3-4):897-931. http://dx.doi.org/10.1016/S0388-0001(96)00054-X

GRAETZ, N. 1985. Teaching EFL students to extract structural information from abstracts. *In:* J.M. ULIJN; A.K. PUGH (eds.), *Reading for professional purposes: Methods and materials in teaching language*, Leuven, Acco, p. 125-135.

JACOBI, D. 1999. *La communication scientifique. Discours, figures, modèles*. Grenoble, Presses Universitaires de Grenoble, 277 p.

KONG, K.C.C. 1998. Are simple business request letters really simple? A comparison of Chinese and English business request letters. *Text*, **18**(1):103-141.

KAPLAN, R.B.; CANTOR, S.; HAGSTROM, C.; KAMHI-STEIN, L.D.; SHIOTANI, Y.; ZIMMERMAN, C.B. 1994. On abstract writing. *Text*, **14**(3):401-426.

L'HOMME, M.C. 1992. *Contribution à l'analyse grammaticale de la langue de spécialité : le mode, le temps et la personne du verbe dans quelques textes scientifiques écrits à vocation pédagogique*. Québec, Canadá. Tesis doctoral. Université Laval, 310 p.

LOFFLER-LAURIAN, A-M. 1983. Typologie des discours scientifiques: Deux approches. *Etudes de Linguistique Appliquée*, **51**:8-20.

MANN, W.C.; TABOADA, M. 2010. *RST Web Site*. Disponible en: http://www.sfu.ca/rst. Acesso em: 08/11/2010.

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

MARCU, D.; CARLSON, L.; WATANABE, M. 2000. The automatic translation of discourse structures. *In:* ANNUAL MEETING OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, I, Seattle, 2000. *Anais...* Seattle, p. 9-17.

MONTOLÍO, E. 2001. *Conectores de la lengua escrita. Contraargumentativos, consecutivos, aditivos y organizadores de la información*, Barcelona, Ariel, 173 p.

MORTUREUX, M-F. 1988. Linguistique et vulgarisation scientifique. *Information sur les sciences sociales*, **24**(4):825-845.

O'DONNELL, M. 2000. RSTTOOL 2.4 – A markup tool for rhetorical structure theory. *In*: INTERNATIONAL NATURAL LANGUAGE GENERATION CONFERENCE, I, Israel. *Anais...* Israel, p. 253-256.

OLIVARES, A. 2006. Divulgación y enfermedades: algunas reflexiones sobre procedimientos de "acercamiento" al gran público. *In:* B. GALLARDO; C. HERNÁNDEZ; V. MORENO (eds.), *Lingüística clínica y neuropsicología cognitiva. Actas del Primer Congreso Nacional de Lingüística Clínica*. Valencia, Universitat de València, p. 96-113.

PÉREZ, L. 2001. *Análisis retórico contrastivo: el resumen lingüístico y médico en inglés y español*. Alicante, España. Tesis doctoral. Universidad de Valladolid, 599 p.

Portolés, J. 2001. *Marcadores del discurso*, Barcelona, Ariel, 160 p.

RAMSAY, G. 2000. Linearity in rhetorical organisation: A comparative cross-cultural analysis of newstext from the People's Republic of China and Australia. *International Journal of Applied Linguistics*, **10**(2):241-258. http://dx.doi.org/10.1111/j.1473-4192.2000.tb00150.x

RAMSAY, G. 2001. What are they getting at? Placement of important ideas in Chinese newstext: Acontrastive analysis with Australian newstext. *Australian Review of Applied Linguistics*, **24**(2):17-34.

REY, J. 1999. Les mécanismes d'introduction d'explications dans les textes scientifiques. *In:* M.A. VEGA; R. MARTÍN-GAITERO (eds.), *Lengua y Cultura. Estudios en torno a la traducción*. Madrid, Universidad Complutense de Madrid, vol. 2, p. 467-472.

REY, J. ; TRICÁS, M. 2006. Stratégies interprétatives et traduction: les introductions et les conclusions dans des textes de semi-vulgarisation scientifique. *Méta. Journal des traducteurs,* **51**:1-19.

SALAGER-MEYER, F. 1991. Medical English abstracts: How well structured are they? *Journal of the American Society for Information Science*, **42**:528-532. http://dx.doi.org/10.1002/(SICI)1097-4571(199108)42:7%3C528::AID-ASI7%3E3.0.CO;2-2

SALKIE, R.; OATES, S.L. 1999. Contrast and concession in French and English. *Languages in Contrast*, **2**(1):27-56. http://dx.doi.org/10.1075/lic.2.1.04sal

SCOTT, D.; DELIN, J.; HARTLEY, A. 1998. Identifying congruent pragmatic relations in procedural texts. *Languages in Contrast*, **1**(1):45-82. http://dx.doi.org/10.1075/lic.1.1.05sco

SWALES, J. 1981. *Aspects of Article Introductions*, Birmingham, The University of Aston, 95 p.

SWALES, J. 1990. *Genre Analysis: English in Academic and Research Settings,* Cambridge, Cambridge University Press, 260 p.

TEUFEL, S.; MOENS, M. 2002. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, **28**:409-445. http://dx.doi.org/10.1162/089120102762671936

TOFILOSKI, M.; BROOKE, J.; TABOADA, M. 2009. A Syntactic and Lexical-Based Discourse Segmenter. *In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 47,* Singapur, 2009. *Anais...* Singapur, p. 77-80.

VAN DIJK, T.A. 1980. *Macro-Structures. An interdisciplinary study of global structures in discourse, cognitions an interaction*, Hillsdale NJ, Erlbaum, 317 p.

VAN DIJK, T.A. 1989. *La ciencia del texto*, Barcelona, Paidós, 310 p.

Submetido em: 31/08/2010
Aceito em: 30/11/2010

Mikel Iruskieta
IXA Group for NLP
Universidad del País Vasco (UPV-EHU)
Bilboko irakasleen Unibertsitate Eskola
Ramón y Cajal 72
48014 Bilbao, Spain

Iria da Cunha
Grup Iulaterm
Institut Universitari de Lingüística Aplicada (IULA)
Universitat Pompeu Fabra (UPF)
C/ Roc Boronal 138, 3ª planta
08018 Barcelona, Spain

ANEXOS

Anexo 1. Lista de relaciones retóricas empleadas en nuestro trabajo.
Appendix 1. List of the rhetorical relations used in our work.

Relación	Tipo
Contraste	N-N
Unión	N-N
Lista	N-N
Secuencia	N-N
Disyunción	N-N
Conjunción	N-N
Fondo	N-S
Circunstancia	N-S
Concesión	N-S
Condición	N-S
Elaboración	N-S
Justificación	N-S
Propósito	N-S
Reformulación	N-S
Resultado	N-S
Resumen	N-S
Evidencia	N-S
Intepretación	N-S
Alternativa	N-S
Antítesis	N-S
Capacitación	N-S
Causa	N-S
Evaluación	N-S
Motivación	N-S
Preparación	N-S
Solución	N-S
Medio	N-S
Condición inversa	N-S
No-condicional	N-S

Anexo 2. Información sobre los textos del corpus analizado.
Appendix 2. Information about the texts of the analyzed corpus.

CORPUS MÉDICO			
Referencia	**Título**	**Autor(es)**	**Año**
Texto 1	Estudio farmacoepidemiologico y farmacoeconómico de la hipertensión arterial	L. C. Abecia	2008
Texto 2	Criterios psicosomáticos de gravedad en oncología	R. Ruiz, A. Aljelani, U. Shelick, U. Usobiaga, J. Muro, J. Bilbao, F. Franco	2007
Texto 3	El fenotipo tumoral "basal like" (c-erb-B2 -, RE - y RP - negativo) define un subgrupo biológicamente muy agresivo de cáncer de mama en estadio postquirúrgico pT1	J. Schneider, A. Tejerina, C. Perea, A. Tejerina R. Lucas, J. Sánchez	2007
Texto 4	Incidencia real de invasión ganglionar de la axila en cáncer de mama T1 en nuestra población	J. Schneider, A. Tejerina, J. Sánchez, J. Lucas	2007
Texto 5	Infección Protésica de Rodilla	O. Sáez-de-Ugarte-Sobrón, I. Gutiérrez-Sánchez, A. Cruchaga-Celada, F. Labayru-Etxebarria, I. Garcia Sánchez, A. Álvarez-González	2008
Texto 6	La Estomatitis Aftosa Recurrente (I): Epidemiología, etiopatogenia y aspectos clínicopatológicos	A. Eguía, R. Saldón, J. M. Aguirre	2003
Texto 7	La cirugía de la bifurcación carotídea en la isquemia cerebral de origen extracraneal: 10 años de experiencia	L. Estallo, A. Barba, L. Rodríguez, S. Gimena, A. G. Alfageme	2000
Texto 8	Manifestaciones infrecuentes de la enfermedad de Whipple. Estudio de cuatro casos	E. Ojeda, A. Cosme, J. Lapaza, J. Torrado, I. Arruabarrena, L. Alzate.	2005
Texto 9	Evolución de las medidas antropométricas del pie infantil. Índices de correlación con otras variables	R. De los Mozos, A. Alfageme, E. Ayerdi	2002
Texto 10	Evolución de las medidas antropométricas del pie infantil. Estudio descriptivo estratificado	R. De los Mozos, A. Alfageme, E. Ayerdi	2002
Texto 11	Evolución de las medidas antropométricas del pie infantil. Estudio descriptivo global	R. De los Mozos Bozalongo, A. Alfageme Cruz, E. Ayerdi Salazar	2003
Texto 12	Atención del ictus y posibilidades de mejora	J. Pérez-de-Arriba, G. Achutegui, L. Epelde, G. Viñegra, JL. Elexpuru.	2005
Texto 13	Morbilidad y tolerancia de la biopsia transrectal ecodirigida en 392 pacientes	J. A. López-Lendoiro, P. Aísa, X. Aguirre, E. Añorbe, M. Paraíso	2002
Texto 14	Tratamiento del pie plano flexible infantil con la técnica de calcáneo-stop	I. Etxebarria-Foronda, I. Garmilla-Iglesias, A. Gay-Vitoria, J. Molano-Muñoz. D. Izal-Miranda, E. Esnal-Baza, A. Ruiz-Sánchez.	2006
Texto 15	Perfil del usuario de la zona ambulatoria del Servicio de Urgencias del Hospital de Galdakao	I. Bengoetxea Martínez	2004
Texto 16	Demencia de rápida progresión y mioclonías	I. Villamil-Cajoto, A, M. J. González-Quintela, V. Villacian-Vicedo	2005
Texto 17	Correlación ecográfica y quirúrgica en las rupturas de grosor completo del manguito rotador de hombro	J. de la Fuente-Ortiz-de-Zárate, J. Kutz-Peyroncelli, J. L. Imizcoz-Barriola	2004

Anexo 2. Continuação.

CORPUS MÉDICO			
Referencia	**Título**	**Autor(es)**	**Año**
Texto 18	Tratamiento quirúrgico de la obesidad mórbida.	I. Díez-del-Val, C. Martínez-Blázquez, V. Sierra-Esteban, J. M. Vitores-López, J. Valencia-Cortejoso	2005
Texto 19	Evolución de los pacientes sometidos a colapsoterapia por tuberculosis pulmonar	K. Abu-Shams, J. Ardanaz, M. Murie, A. Sebastián, G. Tiberio, A. Arteche.	2000
Texto 20	Colonización-infección por Pseudomonas aeruginosa en pacientes con bronquiectasias y EPOC. Aspectos clínicos microbiológicos y evolutivos	J. Garrós Garay, E. Ruiz de Gordejuela, G. Martín Saco, L. Gallego, J. Pérez Escajadillo, F. García Cebrián	2002

CORPUS TERMINOLÓGICO			
Referencia	**Título**	**Autor(es)**	**Año**
Texto 1	Tendencias generales de la normalización en la terminología cientificotécnica de la lengua serbia: análisis crítico de la situación	J. Filipoviçc, J. Filipoviçc	1997
Texto 2	Dimensión social de la normalización terminológica	R. Colomer	1997
Texto 3	La creación de términos para la educación universitaria y para la formación profesional en habla irlandesa	D. Du Bhraonáin, A. Dhubhghaill	1997
Texto 4	La terminología, la traducción y el discurso jurídico desde el punto de vista del euskera	Universidad de Deusto	1997
Texto 5	Metodología del vaciado de terminología vasca de textos legales traducidos	A. Elosegi	1997
Texto 6	El diseño y la gestión de las bases de datos terminológicos: un desafío permanente	J. Bofias, A. Puiggené	1997
Texto 7	Gestión terminológica y proceso cognitivo en lenguas minoritarias	K. Ahmad	1997
Texto 8	El vaciado terminológico automático y su aplicación para el euskera	I. Aldezabal, I. Alegria, X. Artola, N. Ezeiza, R. Urizar	1997
Texto 9	Un constructor terminológico automatizado para el chino	S. C. Lun	1997
Texto 10	Provisión de herramientas terminológicas para la formación profesional en lenguas minoritarias en cuanto al uso y a la enseñanza: proyecto VOCALL	A. Way	1997
Texto 11	Aportes del Tesauro del sistema argentino de informática jurídica a la terminología jurídica	A. Oses	1997

Anexo 3. Análisis cuantitativo de los marcadores detectados en el corpus en español.
Appendix 3. Quantitative analysis of the markers detected into the Spanish corpus.

Relación	Tipo	Marcadores corpus terminológico	T	Marcadores corpus médico	T
Elaboración	N-S	de acuerdo a (1) que (1) y (4) asimismo (1) lo que (1) efectivamente (1) como ejemplo de lo anterior (1) además (1) como (1) también (1)	13	como (2) en consecuencia (1) de ellos (1)	4
Lista	N-N	por un lado (1) y, por otro, (1) y (8) también (1) y por último (1) por otra parte (1)	13	y (16) así mismo (1) así como (3) aparte de (1) igualmente (2)	23
Preparación	N-S	-	0	-	0
Fondo	N-S	-	0	-	0
Resultado	N-S	por lo tanto (2)	2	-	0
Justificación	N-S	como quiera que (1) en la seguridad de que (1) como (1)	3	-	0
Concesión	N-S	aunque (3) si bien (2) de cualquier forma (1)	6	aunque (1)	1
Medio	N-S	-	0	para ello (1) mediante (1)	2
Propósito	N-S	con este fin (1) el objetivo es (1) a fin de (1) con tal fin (1) el objetivo de nuestro proyecto es (1) para (1)	6	para (1) con el fin de (1)	2
Secuencia	N-N	seguidamente (2) a continuación (1)	3	-	0
Antítesis	N-S	de todas formas (1) pero (2) de todos modos (1)	4	pero (1)	1
Causa	N-S	ya que (2) habida cuenta de que (1) por consiguiente (1)	4	debido a (1) por lo que (2)	3
Condición	N-S	si (2) siempre que (1)	3	-	0
Motivación	N-S	de manera que (1) con esto en mente (1)	2	por lo que (1)	1
Interpretación	N-S	por lo que (1)	1	lo que (1)	1

Anexo 3. Continuação.

Relación	Tipo	Marcadores corpus terminológico	T	Marcadores corpus médico	T
Circunstancia	N-S	desde que (1) mientras (1)	2	a la hora de (1)	1
Evaluación	N-S	-	0	-	0
Reformulación	N-S	es decir (1)	1	-	0
Resumen	N-S	tal y como hemos indicado más arriba (1)	1	-	0
Solución	N-S	-	0	-	0
Disyunción	N-N	-	0	-	0
Capacitación	N-S	-	0	-	0
Conjunción	N-S	-	0	-	0
Evidencia	N-S	-	0	como (1)	1
Alternativa	N-S	-	0	-	0
Unión	N-S	y (2)	2	-	0
Contraste	N-N	-	0	mientras que (2) y (2) o (1) por lo demás (1) frente a (1) por el contrario (1)	8
Cond. inversa	N-S	-	0	-	0
No-condicional	N-S	-	0	-	0
Total			**59**		**48**
Total N-N			**16**		**32**
Total N-S			**43**		**16**

Anexo 4. Análisis cuantitativo de los marcadores detectados en el corpus en euskera.
Appendix 4. Quantitative analysis of the markers detected into the Basque corpus.

Relación	Tipo	Marcadores corpus terminológico	T	Marcadores corpus médico	T
Elaboración	N-S	horien arabera (1) eta (2) hala da (1) zelanbait (1) era berean (1) aldi berean (1) eta batez ere (1) batez ere (2) horren adibide gisa (1)	11	horien artean horren ondorioz eta (5) horien artetik berriz beraz gainera (2) -larik era berean (2) hala nola (2)	17
Lista	N-N	alde batetik eta bestetik (1) eta (4) ere (1) -gatik (1) -lako (1) beste aldetik (1)	9	horrez gain (1) eta (10) ere (3) era berean (6)	20
Preparación	N-S	-	0	-	0
Fondo	N-S	-z (1)	1	-	0
Resultado	N-S	honela (1) -nez (1)	2	eta (1) horien artean (1) horrenbestez (1)	3
Justificación	N-S	gainera (1)	1	izan ere (1) horrela (1) horrela gauzak (1) -lako (1) bait- (1)	5
Concesión	N-S	nahiz eta (1) hala ere (3) arren (1) baina (2) ba- ere (1)	8	-n arren (1) oraindik ere (1) baina (1)	3
Medio	N-S	horretarako (1) xede hori iristeko (2)	2	besteak beste (1) -z (1)	2
Propósito	N-S	-lakoan (1) horretarako (1) subjuntivo (2) proiektuaren helburua (1)	5	-teko asmoz (1) -teko helburuarekin (1) hau dela eta (1) -tzea (1)	4
Secuencia	N-N	batetik (1) hurrengo eta, amaitu orduko (1) eta (2) ondoren (2) gero (1) ondorioa (1)	8	-	0
Antítesis	N-S	-	0	baina (1)	1
Causa	N-S	izan ere (1) eta (3) bait- (2) -nez (1) horren ondorioz (2) -nez gero (1)	10	-	0

Anexo 4. Continuação.

Relación	Tipo	Marcadores corpus terminológico	T	Marcadores corpus médico	T
Condición	N-S	-ez gero (1) behintzat (1) ba- (2)	4	ba- (2)	2
Motivación	N-S	-nez (1) -teko (1)	2	-	0
Interpretación	N-S	-	0	beraz (1) ondorioz (1) eta (1)	3
Circunstancia	N-S	-n bitartean (1)	1	-tzean (1)	1
Evaluación	N-S	-	0	-	0
Reformulación	N-S	are gehiago (1) hau da (1) gorago esan bezala (1)	3	-	0
Resumen	N-S	-	0	-	0
Solución	N-S	-	0	-	0
Disyunción	N-N	-	0	edo (1)	1
Capacitación	N-S	-	0	-	0
Conjunción	N-S	eta (6) eta gero (1) era berean (1) ere (1) beste aldetik (1)	10	-	0
Evidencia	N-S	beraz (1)	1	hori adierazten du (1)	1
Alternativa	N-S	-	0	-	0
Unión	N-S	-	0	-	0
Contraste	N-N	baina (1)	1	bestalde (2) aldiz (2) eta berriz (1) berriz (1) gainerakoan (1)	7
Cond. inversa	N-S	-	0	-	0
No condicional	N-S	-	0	-	0
Total			**79**		**70**
Total N-N			**28**		**28**
Total N-S			**51**		**42**

Anexo 5: Traducción euskera-español de los marcadores incluidos en el Anexo 4*.
Appendix 5: Basque-Spanish translation of the markers included into the Appendix 4.

	Marcadores en español	Marcadores en euskera
1	a continuación / y por último	hurrengo eta, amaitu orduko
2	a la hora de	-t(z)eko orduan-
3	además	gaitera
4	al menos	behintzat
5	aparte de	horrez gain
6	así las cosas	horrela gauzak
7	así mismo	era berean
8	asimismo	baita ere
9	aún más	are gehiago
10	aún todavía	oraindik ere
11	aunque	ba-... ere
12	aunque	nahiz eta
13	como	-(e)nez
14	como ejemplo de lo anterior	horren adibide gisa
15	como quiera que	-z
16	con tal fin / a fin de / con este fin	-t(z)eko helburuarekin
17	de acuerdo a	horien arabera
18	de alguna forma	zelanbait
19	de ellos	horien artean
20	de ellos	horien artetik
21	de ese modo	horrela
22	de hecho	izan ere
23	de manera que	honela,
24	de manera que	-t(z)eko
25	debido a	hau dela eta
26	debido a / que	-(e)lako
27	efectivamente	hala da
28	el objetivo de nuestro proyecto es	Proiektuaren helburua
29	el objetivo es	helburua da
30	en conclusión	horrenbestez
31	en conclusión	ondorioa,
32	en consecuencia	beraz
33	en la seguridad de que	-(e)lakoan
34	entre otros	besteak beste
35	es decir	hau da
36	frente a	ostera
37	habida cuenta de que	-(e)nez
38	igualmente	halaber
39	lo que / por lo que	-(e)la eta
40	más tarde	gero
41	mediante	-(e)n bidez
42	mientras	bitartean

Anexo 5: Continuação.

	Marcadores en español	Marcadores en euskera
43	mientras que	aldiz
44	o	edo
45	para	horretarako
46	para ello	horretarako
47	para poder alcanzar es objetivo	xede hori iristeko
48	pero	baina
49	por consiguiente	horren ondorioz
50	por ejemplo	hala nola
51	por el contrario	berriz
52	por lo demás	bestela
53	por lo tanto	ondorioz
54	por un lado	batetik
55	por un lado y por otro	alde batetik eta bestetik
56	porque	-gatik
57	seguidamente	ondoren
58	si	ba-
59	si bien	arren
60	si es que	-(e)nez gero
61	siempre que	baldin eta
62	simultáneamente	aldi berean
63	sin embargo	bestalde
64	sin embargo / de cualquier forma / de todas formas / de todos modos	hala ere
65	sobre todo	batez ere
66	tal y como hemos indicado más arriba	gorago esan bezala
67	tambien	ere
68	y	eta
69	y + en cambio	eta + berriz
70	y despues	eta gero
71	y, por otro,	beste aldetik

* Las traducciones han sido extraídas de los textos paralelos, salvo en algunas excepciones.