Bases para la implementación de un segmentador discursivo para el euskera

Mikel Iruskieta[1], Arantza Diaz de Ilarraza[2], Mikel Lersundi[3]

IXA group for NLP. Faculty of informatics. University of the Basque Country
Post code 10018 – Donostia – Basque Country - Spain

[1]Department of Didactics of Language and Literature
[2]Department of Computer Science
[3]Department of Basque Philology

{mikel.iruskieta, a.diazdeilarraza, mikel.lersundi}@ehu.es

Abstract. *In this paper we study how to adapt an automatic clause parser to discourse segmentation task. Considering a manually tagged corpus according to Rhetorical Structure Theory (RST), we have processed it with an automatic clause parser and the results were studied by comparing the agreement between both annotation systems: automatic and manual. As a result of this comparison we indicate where the intersection among the automatic clause segmentation and discursive segmentation is.*
Keywords. *Discourse segmentation; Rhetorical Structure Theory, parser.*

Resumen. *Presentamos un estudio para adaptar el segmentador automático de cláusulas y oraciones de carácter general para el euskera a la tarea de segmentación discursiva. Partiendo de un corpus anotado manualmente según la Rethorical Structure Theory (RST), hemos procesado el texto de manera automática por medio del segmentador automático y hemos estudiado los resultados comparando las coincidencias y desacuerdos entre la anotación automática y la manual. Los resultados de esta comparación señalan los criterios comunes para adaptar el segmentador a tareas discursivas.*
Palabras clave. *Segmentación discursiva, Teoría de la Estructura Retórica, segmentador.*

1. Introducción

En este artículo presentamos un estudio para adaptar el segmentador de cláusulas y oraciones de carácter general que disponemos para el euskera a la tarea de segmentación discursiva. El segmentador que analizamos ha sido utilizado, en concreto, para tareas de corrección de puntuación en textos (Arrieta 2010) y está implementado mediante la combinación de gramáticas basadas en reglas y técnicas de aprendizaje automático.

En este trabajo trataremos de responder a las siguientes cuestiones: ¿es adecuado abordar la tarea de la segmentación discursiva partiendo de un segmentador de cláusulas y oraciones de carácter general?, ¿cuáles son los criterios comunes entre la segmentación sintáctica y la segmentación discursiva?, y ¿cuándo podemos concluir que es aceptable la segmentación automática discursiva?

Aplicaciones avanzadas, tales como la búsqueda de información basada en conocimiento semántico, la elaboración automática de resúmenes o la traducción automática, precisan herramientas sofisticadas de procesamiento del lenguaje que, a su vez, necesitan basarse en el conocimiento presente en el corpus. Por ello, y para poder llevar a cabo este tipo de aplicaciones, es necesario contar con corpus de referencia

Anais do III Workshop "A RST e os Estudos do Texto", páginas 18–29, Cuiabá, MT, Brasil, Outubro 24–26, 2011.

etiquetados a diferentes niveles lingüísticos: fonético, morfológico, sintáctico o discursivo.

El etiquetado de corpus de referencia en cualquiera de los niveles de análisis lingüístico tiene como primer paso la segmentación. Ésta consiste en identificar y marcar las unidades básicas a considerar en cada nivel lingüístico de análisis, para después determinar las relaciones entre dichas unidades. La identificación de fonemas y su anotación en los corpora es una tarea necesaria para el tratamiento del habla, como es la identificación por un lado de lexemas y morfemas, y por otro de sintagmas y dependencias son necesarias en el etiquetado de corpora a nivel morfológico y sintáctico. También es ineludible la segmentación a nivel discursivo para identificar la estructura relacional de un texto. Este trabajo trata precisamente de la segmentación de este último nivel: el nivel discursivo.

Atendiendo a la granularidad con la que se establece la unidad de discurso, encontramos en la literatura diferentes propuestas para la segmentación discursiva. Las propuestas varían según la aproximación teórica usada y según la finalidad para la que se realiza el trabajo de etiquetado. En general podemos distinguir entre dos niveles en la segmentación discursiva: segmentación de nivel alto y segmentación de nivel bajo. En esta última, a su vez se distinguen dos subniveles: intra-oracional (mayor granularidad) e inter-oracional (menor granularidad). Por ejemplo la segmentación intra-oracional donde se establecen unidades de discurso a nivel de cláusula es utilizada en Marcu (2000) para tareas de resumen automático. La segmentación de alto nivel donde se establecen pasajes o párrafos es utilizada en tareas de recuperación de la información (Girill 1991) o detección de cambios de tópico (Hearst 1997). En este trabajo abordaremos la segmentación intra-oracional, ya que nuestro objetivo es la anotación de corpus válidos para una amplia variedad de aplicaciones.

En la literatura se referencian segmentadores de discurso "independientes de lenguaje" (Kiss y Strunk 2006) que detectan segmentos únicamente a nivel inter-oracional. En el corpus sobre el que hemos trabajado los segmentos a nivel intra-oracional suponen alrededor de un 9%. En la actualidad conocemos herramientas de segmentación discursivas de nivel bajo para inglés, portugués y español (Tofiloski, Brooke y Taboada 2009, Pardo 2006, da Cunha*, et al* 2010). Hasta el momento no existe una herramienta de dichas características en euskera y este es el objetivo que nos proponemos a corto plazo. Este trabajo supone un paso importante en la consecución en ese objetivo.

2. Estado del arte: teorías y corpora

Existen diferentes teorías discursivas que formalizan la estructura referencial; cada una de estas teorías proporciona corpora anotados según sus criterios: i) Segmented Discourse Representation Theory (SDRT) (Asher y Lascarides 2003); ii) Discourse-Lexicalized Tree Adjoining Grammar (D-LTAG) (Webber*, et al* 2003); y, iii) Rhetorical Structure Theory[i] (RST) (Mann y Thompson 1987). Esta última teoría describe la coherencia y relación entre fragmentos textuales haciendo corresponder la idea de nuclearidad, o importancia de un fragmento del discurso, con el efecto que produce en el lector la presentación de dicha relación. Cuenta con varios corpus para diferentes lenguas: i) para el inglés, un corpus de 385 textos periodísticos (Carlson, Okurowski y

Marcu 2002) y otro de 65 textos de géneros diferentes (Taboada y Renkema 2011); ii) para el español un corpus de 267 textos (da Cunha, Torres-Moreno y Sierra 2011); iii) para el portugués el corpus TCC de 100 textos científicos (Pardo y Nunes 2006), y iv) para el alemán el corpus PCC de 170 textos etiquetados (Stede 2004). Existen segmentadores discursivos para el inglés (Marcu 2000, Tofiloski, Brooke y Taboada 2009), para el portugués (Pardo y Nunes 2008) y el español (da Cunha*, et al* 2010)[ii]. La RST ha sido implementada para diversas aplicaciones de PLN según Taboada y Mann (2006a).

El marco teórico sobre el que desarrollamos este estudio empírico es la RST. Según esta teoría, las relaciones que se establecen entre los segmentos del texto pueden ser paratácticas (N-N)[iii], cuando se establece la relación entre fragmentos con el mismo grado de importancia en la intención del autor (LISTA, CONTRASTE, DISYUNCIÓN…), o hipotácticas (N-S), cuando se establece una relación entre una unidad menos importante con otra más importante en cuanto a la intención del autor (ELABORACIÓN, MÉTODO, PREPARACIÓN, CONCESIÓN, CAUSA, RESULTADO…). Las relaciones se definen en base a las restricciones presentes entre el núcleo y satélite, y describiendo el efecto que crea en el lector.

El corpus sobre el que hemos realizado el estudio es un corpus de resúmenes de artículos médicos extraídos de la Gaceta Médica de Bilbao[iv], que contiene todos los resúmenes de artículos en euskera desde sus inicios en el año 2000 hasta el 2008. El corpus está compuesto por 20 documentos y contiene 273 unidades elementales de discurso (EDU); a nivel intra-oracional cada EDU tiene como media unas 11 palabras, y el corpus tiene 3.024 palabras. Este corpus ha sido utilizado en trabajos anteriores (da Cunha y Iruskieta 2010) donde se sugiere que se pueden detectar estrategias de traducción mediante la comparación de árboles retóricos en idiomas diferentes. La anotación de este corpus está disponible tanto en español (da Cunha, Torres-Moreno y Sierra 2011) como en euskera[v].

Aunque en la RST existen diferentes propuestas para la segmentación de textos, el corpus en el que nos basamos se ha segmentado siguiendo la definición original de unidad básica de Mann y Thompson (1987) que dice fundamentarse en una clasificación teórica neutral en la que las unidades debieran caracterizarse por una integridad funcional independiente.

3. Segmentación manual y automática. Comparación

Para determinar si el segmentador automático es un buen punto de partida en la construcción de un segmentador discursivo, vamos a comparar el resultado del segmentador de cláusulas y oraciones con nuestra anotación discursiva manual que sigue la segmentación original de la RST y establecer criterios comunes para definir reglas básicas de implementación válidas en el marco de la RST.

En lo referente a la segmentación manual, y, tras un proceso escalonado para establecer los criterios de segmentación (Iruskieta, Díaz de Ilarraza y Lersundi En prensa), se han fijado las siguientes reglas de segmentación a nivel inter-oracional e intra-oracional: i) en el nivel inter-oracional se van a considerar unidades de discurso aquellas oraciones con verbo conjugado no subordinadas[vi], y ii) en el nivel intra-oracional se consideran unidades de discurso oraciones con verbo (tanto conjugado

como no conjugado). Los complementos verbales no se consideran unidad del discurso aunque posean formas verbales (por ejemplo, complementos de verbos declarativos).

En referencia a la segmentación automática, nuestro sistema para la segmentación del corpus médico utiliza el sistema descrito en (Alegria*, et al* 2008) que identifica cláusulas mediante la combinación de gramáticas basadas en reglas y técnicas de aprendizaje automático. Las reglas establecen los puntos donde finalizan las oraciones y mediante las técnicas de aprendizaje automático se reconocen el comienzo y final de las estructuras sintácticas parciales basándose en la información lingüística asociada a cada palabra de la oración (Carreras 2005). La información se ha obtenido tras la aplicación de la siguiente secuencia de tratamientos lingüísticos:

1. Análisis morfo-sintáctico (MORPHEUS[vii] (Aduriz*, et al* 1998)). Proceso por el cual se establece la segmentación de cada palabra, su categoría, subcategoría y otras características lingüísticas tales como caso, número, etc. El principal problema de este paso de análisis es la gran cantidad de análisis asociados a cada palabra, ya que el análisis de la palabra se realiza sin tomar en cuenta el contexto en el que se encuentra.

2. Lematización e identificación de funciones sintácticas. Estos dos procesos se realizan en secuencia mediante la aplicación EUSTAGGER[viii] (Aduriz*, et al* 2003). La principal tarea del lematizador es resolver la ambigüedad que resulta del proceso de análisis morfo-sintáctico tratando de dar un único análisis para cada palabra de la frase basándose en la información contextual. La identificación de funciones sintácticas se realiza mediante reglas basadas en conocimiento lingüístico que siguen el formalismo establecido en las gramáticas de restricciones (Karlsson*, et al* 1995).

3. Identificación de unidades multi-palabra cuyo objetivo es determinar las unidades que se componen de dos o más palabras, considerando sólo los casos en que estas asociaciones de palabras sean siempre fijas.

4. Identificación de entidades nombradas (EIHERA[ix] (Alegria*, et al* 2003)).

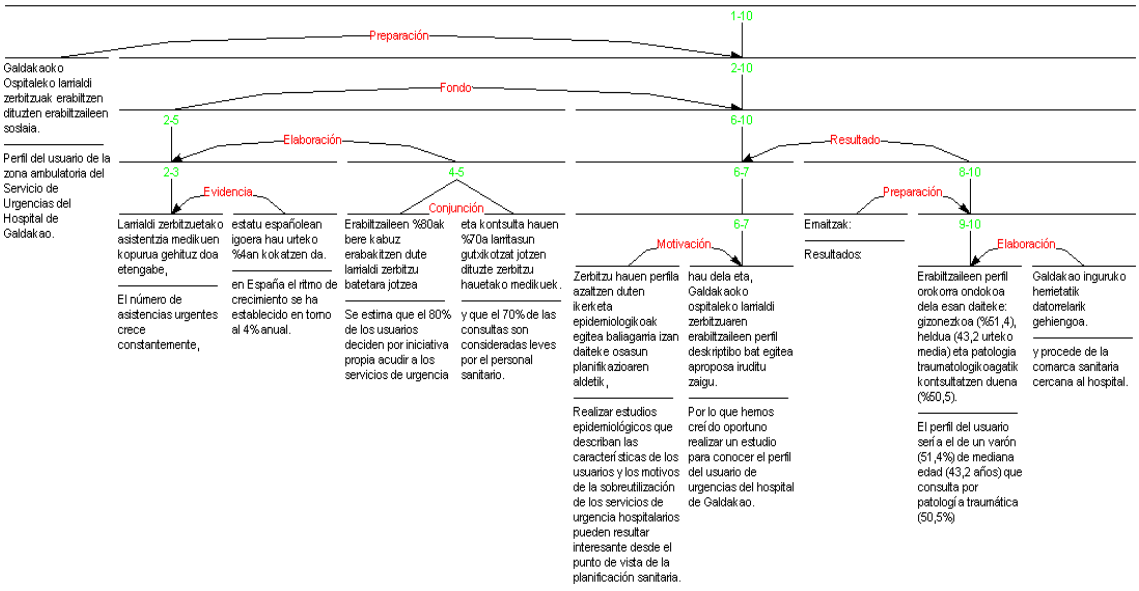

Figura 1. Árbol retórico (GMB_04_01)

En la Tabla 1 se presenta de forma gráfica la segmentación manual y automática[x] del ejemplo (1) tomado del texto del corpus representado en la Figura 1.

(1) a. <[<*Erabiltzaileen %80ak bere kabuz erabakitzen dute*> <*larrialdi zerbitzu batetara jotzea*>] [<*eta kontsulta hauen %70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek.*>]> GMB_04_01

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

EDUs en segmentación manual		**EDUs en segmentación automática**		
M1	**M2**	**A1**	**A2**	**A3**
Erabiltzaileen %80ak bere kabuz erabakitzen dute larrialdi zerbitzu batetara jotzea	*eta kontsulta hauen %70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek.*	*Erabiltzaileen %80ak bere kabuz erabakitzen dute*	*larrialdi zerbitzu batetara jotzea*	*eta kontsulta hauen %70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek.*
Se estima que el 80% de los usuarios deciden por iniciativa propia acudir a los servicios de urgencia	y que el 70% de las consultas son consideradas leves por el personal sanitario.	Se estima que el 80% de los usuarios deciden por iniciativa propia	acudir a los servicios de urgencia	y que el 70% de las consultas son consideradas leves por el personal sanitario.

Tabla 1. Comparación segmentaciones (fragmento de GMB_04_01)

Como hemos comentado la segmentación automática tiene un componente basado en reglas mediante las que se establece la identificación de limites clausales y oracionales. Presentamos en la Tabla 2, a modo de ejemplo, dos de las reglas que se aplicarían para identificar los limites clausales en el texto del ejemplo (1).

N°	**Explicación de la regla**
11	`MAP (}MUGA) TARGET (ADL) IF (1 (LOT)+(JNT))` `(NOT 1 ("baita")OR("ezta"));`

68	`MAP (}MUGA) TARGET (ADIZE) IF (0 (DEK)) (NOT 1 PUNTUAZIOA)`
	`(NOT 1 ("aritu")+(ADOIN))`
	`(NOT -2 ("aritu")+(ADOIN));`

Tabla 2. Reglas de segmentación utilizadas en el ejemplo 1

La regla 11 asigna la marca de fin de segmento de A1 tras el verbo auxiliar (ADL) (*erabakitzen dute)* 'deciden', si y solamente si: i) viene seguido un conector (LOT) y que es a su vez conjunción coordinante (JNT) y ii) inmediatamente a la derecha del auxiliar no están las palabras *baita* 'también' y *ezta* 'tampoco'.[xi]

La regla 68 asigna la frontera de A2 a la nominalización (ADIZE) *jotzea* 'pegar', si y solamente si la nominalización posee alguna marca de declinación (DEK) y i) no tiene signos de puntuación a su derecha, ii) no tiene el verbo *aritu* 'ocuparse'[xii] más una forma verbal sin terminación aspectual (ADOIN) a su derecha o iii) a una distancia de dos palabras a la izquierda. Esta regla y el final del segmento anterior A1 son suficientes para determinar el segmento A2.

Entre la segmentación manual y automática hay diferencias de granularidad que indican que la segmentación automática es más fina que la manual, ya que se consideran criterios más relacionados con la función sintáctica que cumplen las oraciones (esto no se ajusta a lo establecido por nuestras guías de anotación previamente definidas). Por ejemplo, la segmentación automática considera la nominalización *jotzea* 'acudir' como EDU y la segmentación manual lo descarta por considerarse un complemento verbal y, por tanto, no considerar que exista una relación RST.

Las demás reglas de la gramática, al igual que las reglas explicadas que son utilizadas por el segmentador, determinan únicamente el final de cada segmento. Esto no es problema cuando al finalizar un segmento empieza otro, tal como sucede en el ejemplo (1); pero para detectar, además de estos segmentos en secuencia, segmentos subsumidos en otros (como en el ejemplo (2) donde la unidad 3 de la Figura 2, una cláusula adverbial de modo que se enlaza, en este caso, con la relación de MÉTODO, está subsumida dentro de otra unidad formalizada por la construcción SAME-UNIT) hemos utilizado técnicas de aprendizaje automático. Este problema es más crítico en un corrector de signos de puntuación. En la segmentación intra-oracional, sin embargo, las unidades que rompen una EDU no son tan abundantes; en este corpus dichas construcciones constituyen únicamente el 0,03% de todas las unidades.

(2) a. <[<*Ikerketa berriek,*} [*"microarrays" teknika erabiliz,*>] {<*pronostiko txarra duen*> *bularreko minbiziaren azpitalde bat hauteman dute*.]> GMB_07_02

b. <[<Estudios recientes} [utilizando la técnica de "microarrays">] {<han identificado un subgrupo de cánceres de mama <con pésimo pronóstico.>]>

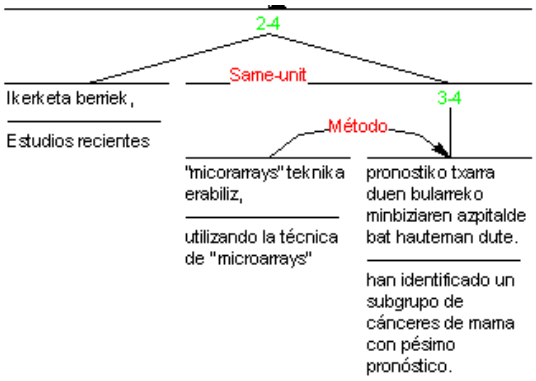

Figura 2. Árbol retórico del ejemplo 2

4. Resultados y evaluación

Actualmente podemos encontrar herramientas de segmentación automática con un *F-measure* en torno a 80%: *SLSeg* en inglés obtiene un 79% de *F-measure* (Tofiloski, Brooke y Taboada 2009) y *DiSeg* en español un 80% (da Cunha*, et al* 2010). Aunque los datos del segmentador automático general que utilizamos son bajos, un *F-measure* de 57,81%, se detectan la mayoría de EDUs; parte de los segmentos (S) que no se detectan se debe a que algunos segmentos intra-oracionales se formalizan de modo diferente, como cláusulas adverbiales y coordinaciones de EDUs. En la Tabla 3 se presentan los acuerdos obtenidos sobre las marcas de inicio de segmento (2ª fila: <S), final de segmento (3ª fila S>), donde el acuerdo es mayor, y la comparación entre las marcas de principio y final de segmento automático y marcas de EDUs (3ª fila: EDU), donde el acuerdo baja considerablemente. La cobertura que mide el grado de marcas automáticas que coinciden con las manuales señala que se han detectado la mayoría de las EDUs de un modo más que aceptable el final de cada segmento. Sin embargo la precisión, que mide el grado de marcas correctas de todas las marcas puestas por el segmentador, disminuye de modo considerable lo que indica una granularidad mayor del segmentador.

	Automático	Manual	Acuerdo	Cobertura	Precisión	F-measure
<S	450	273	223	81,68%	49,56%	61,69%
S>	450	273	242	88,64%	53,78%	66,94%
EDU	450	273	209	76,56%	46,44%	57,81%

Tabla 3. Evaluación del segmentador

Hemos hecho un estudio en detalle para ver por qué no coinciden los segmentos marcados automáticamente con los anotados manualmente y hemos detectado dos fenómenos:

i) Sobre-segmentación: el segmentador automático identifica más segmentos de los anotados manualmente. El segmento A2 *larrialdi zerbitzu batetara jotzea* 'acudir a los servicios de urgencia' del segmentador no se ha considerado en el modo manual, ya que el verbo nominalizado es parte en un sintagma nominal y su relación es puramente sintáctica con referencia al segmento A1.

ii) Falta de segmentación: el segmentador automático no detecta algunos segmentos o no formaliza adecuadamente una EDU con ambas marcas de inicio y final adecuadamente. El segmentador al establecer el segmento M1 *erabiltzaileen %80ak bere kabuz erabakitzen dute larrialdi zerbitzu batetara jotzea* 'se estima que el 80% de los usuarios acuden a los servicios de urgencia por iniciativa propia' en varios segmentos A1 y A2, no formaliza de manera adecuada dicho segmento, es decir que las marcas de inicio y final no coinciden con las de una EDU. En otros casos la falta de segmentación es debida a diferentes modos de formalización en la segmentación automática y manual.

La Tabla 4 y la Tabla 5 muestran numéricamente la frecuencia de aparición de estos dos fenómenos: sobre-segmentación y falta de segmentación, respectivamente. Explicaremos brevemente los casos en que se dan estos fenómenos e intercalaremos algún ejemplo para dar una mayor claridad a la explicación.

En cuanto al fenómeno de la sobre-segmentación hemos identificado los siguientes casos: i) la segmentación automática ha detectado una oración principal, pero no incluye todas las palabras incluidas en la segmentación manual (Oración incompleta): ejemplo (3). El primer segmento automático es adecuado, porque coincide con el segmento manual, pero el segundo segmento automático, no considerado en la segmentación manual, recoge sólo en parte la oración principal; ii) la segmentación automática no formaliza adecuadamente y agrupa varias EDUs, por ejemplo, cuando establece como una EDU construcciones que componen más de una unidad, (Composición de EDUs); el ejemplo (4) se debe a una diferencia de formalización porque la marca de inicio de la segunda EDU no se ha colocado en la posición del segmento manual donde finaliza la oración subordinada adverbial sino que se ha colocado a su inicio, segmentando de este modo toda la oración compuesta[xiii]; iii) los segmentos corresponden a complementos de verbo (complementos, oraciones interrogativas indirectas, nominalizaciones…) y/o modificadores de sintagmas nominales (oraciones de relativo) que no consideramos EDUs en la segmentación manual (Complemento); en el ejemplo (5) observamos una oración relativa; iv) el segmentador identifica como unidad la coordinación de varios complementos o elementos coordinados sintácticamente que no constituyen EDUs (Coordinación de complementos), y finalmente v) cláusulas que no se han considerado

EDUs en la segmentación manual y se han segmentado automáticamente debido a la puntuación (Puntuación).

Oración incompleta	Composición de EDUs	Complemento	Coordinación	Puntuación	Total
13	26	87	74	38	238
5,46%	10,92%	36,55%	31,09%	15,97%	100,00%

Tabla 4. EDUs sobre-segmentados

(3) a. <[1996ko urtarriletik 1996ko ekainera arte, kolapsoterapia hartzen duten 30 gaixo, <batez beste 70.8 ±17 urtekoak (60-83 urte), aztertu ditugu guztira.>]> GMB_00_01

b. <[Desde Enero de 1996 hasta Junio de 1996 <hemos revisado a un total de 30 pacientes con colapsoterapia, con 70.8±17 años (60-83 años) de edad media.>]>[xiv]

(4) a. <[<Prebentzio metodoen eta artroplastiako teknika modernoen laguntzaz horrelako kasuak murriztu diren arren,>] [infekzio hori sendatzea erronka bat da oraindik ere.]> GMB_08_02

b. <[<Aunque su incidencia ha disminuido a lo largo de los años gracias a la evolución de los métodos de prevención y a las técnicas de artroplastia modernas,>] [su tratamiento sigue siendo un reto.]>

(5) a. <[<eta gaur egunera arte deskribatu diren adibideetan daukaten> maiztasuna alderatu da.]>GMB_05_03

b. <[y se compara su frecuencia <entre las series más numerosas de la literatura descritas hasta la actualidad.>]>

En cuanto al fenómeno de la falta de segmentación hemos identificado estos otros casos: i) una oración principal no detectada (Oración principal); en el ejemplo (6) la segmentación automática formaliza de forma diferente el primer segmento, ya que su cierre se introduce al final de la oración y no antes del segundo segmento, por lo que no coinciden las marcas de inicio y final de ambas segmentaciones; ii) no se detectan cláusulas adverbiales (Cláusula adverbial); iii) no se formalizan adecuadamente las unidades por separado que están coordinadas (Coordinación), y finalmente iv) EDUs que no se segmentan de modo adecuado debido a la puntuación (Puntuación).

Oración principal	Cláusula adverbial	Coordinación	Puntuación	Total
20	20	7	17	64
31,25%	31,25%	10,94%	26,56%	100,00%

Tabla 5. EDUs falta de segmentar

(6) a. <[Ultzera mingarri batzuk bezala agertzen da,] [<tamainu, kokapena eta iraunkortasuna aldakorra izanik.>]> GMB_03_01

b. <[Se caracteriza por la aparición de úlceras dolorosas] [<siendo de tamaño, localización y duración variable.>]>

Por último presentamos las unidades discursivas detectadas correctamente por el segmentador automático (Tabla 6).

Únicamente principal	Principal con subordinación	Cláusula adverbial	Yuxtap. o coordinación	Puntuación	Título	Total
64	57	18	51	6	13	**209**
30,62%	27,27%	8,61%	24,40%	2,87%	6,22%	**100,00%**

Tabla 6. EDU detectados correctamente

La comparación realizada entre la anotación de segmentos realizada automáticamente y la manual nos señala cómo adaptar la herramienta automática a la segmentación de discurso. En la Tabla 7 presentamos las conclusiones de la comparación.

	Forma lingüística	Segmentador general	Segmentador discursivo
Principios generales	Oración o cláusula verbal	sí	sí
	Same-unit construcción	sí	sí
Subordinación	Cláusulas adverbiales	sí	sí
	Complementos con clausulas verbales	sí	no
	Oración interrogativa indirecta	sí	no
	Cláusulas comparativas	ssi cláusulas verbales	no
	Nominalización	sí	no
	Clausulas de relativo	sí	no
Coordinación y/o yuxtaposición	de cláusulas verbales que difieren en un argumento	sí	sí
	de cláusulas verbales sin argumentos propios	sí	no
	de clausulas adverbiales	sí	sí
	de clausulas no-adverbiales	sí	no
	de cláusulas no verbal con marcador	sí	no
	Locuciones con función relacional	sí	no
Puntuación	Cláusulas verbales parentéticas	sí	sí
	Cláusulas no-verbales parentéticas	no	no
	Cláusulas de aposición	no	no
	Punto oracional con o sin verbo	sí	sí
	Dos puntos	sí	ssi EDU después
	Punto y coma	sí	ssi EDU después

Tabla 7. Criterios generales de adaptabilidad

5. Conclusiones y trabajo futuro

El estudio demuestra que aunque el porcentaje de EDUs segmentados correctamente (precisión en Tabla 3) por el segmentador automático es bajo y, por ello dicha segmentación no es la adecuada para la posterior anotación retórica en el marco de la RST; el método seguido para lograr un segmentador discursivo automático es un buen punto de partida, ya que el segmentador ha segmentado adecuadamente la mayoría de EDUs (cobertura Tabla 3). Para lograr ese objetivo, hemos detectado de manera precisa en qué situaciones no coinciden las marcas identificadas por el segmentador automático y qué nuevos criterios de segmentación debemos incorporar en el segmentador

automático, lo que supone un primer paso en la consecución de segmentador de discurso automático válido para la RST.

Teniendo en cuenta que el segmentador automático del que partimos está basado en reglas lingüísticas y algoritmos de aprendizaje automático, en el futuro nos proponemos realizar la tarea de adaptación de algunas de las reglas y adición de nuevas reglas del segmentador automático. Además tendremos que llevar a cabo el reentrenamiento del componente basado en aprendizaje automático tomando como base el corpus etiquetado que se obtendría al aplicar la gramática "adaptada" a un conjunto de textos (corpus de entrenamiento).

Agradecimientos

Este trabajo ha sido realizado en el marco de los siguientes proyectos: Grupo IXA, Grupo consolidado 2010-2015 (IT344-10) (Gobierno Vasco); KNOW2: Tecnologías de comprensión del lenguaje para el acceso multilingüe a la información orientada a dominios (TIN2009-14715-C04-01) (MICINN); Hibrido Sint: analizadores sintácticos basados en reglas y estadísticos. Integración en una plataforma para gestión de corpus basada en estándares XML (TIN2010-20218) (MICINN); Desarrollo de un entorno para extraer terminología y neología a partir de corpus etiquetados lingüísticamente GARATERM2 (US10/01).

Referencias

Aduriz, I., E. Agirre, I. Aldezabal, I. Alegria, O. Ansa, X. Arregi, J. Arriola, X. Artola, A. Díaz de Ilarraza y N. Ezeiza, 1998. A framework for the automatic processing of Basque. En *Proceedings of the First International Conference on Language Resources and Evaluation*.

Aduriz, I., I. Aldezabal, I. Alegria, J. Arriola, A. Díaz de Ilarraza, N. Ezeiza y K. Gojenola, 2003. Finite state applications for basque. En *EACL 2003 Workshop on Finite-State Methods in Natural Language Processing*.

Alegria, I., B. Arrieta, X. Carreras, A. Díaz de Ilarraza y L. Uria, 2008. Chunk and clause identification for basque by filtering and ranking with perceptrons. *Procesamiento del lenguaje natural,* (41): 5-12.

Alegria, I., I. Balza, N. Ezeiza, I. Fernandez y R. Urizar, 2003. Named entity recognition and classification for texts in basque. En *II Jornadas de Tratamiento y Recuperación de Información*, 1-8.

Arrieta, B., 2010. *Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean.* Tesis doctoral. EHU: Euskal Herriko Unibertsitatea.

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

Carlson, L., M.E. Okurowski, D. Marcu, 2002. RST Discourse Treebank[Corpus]. *Linguistic Data Consortium.*

Carreras, X., 2005. *Learning and inference in phrase recognition: a filtering-ranking architecture using perceptron.* Tesis doctoral. Polytechnic University of Catalunya.

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

da Cunha, I., E. SanJuan, J.M. Torres-Moreno, M. Lloberes y I. Castellón, 2010. Discourse segmentation for Spanish based on shallow parsing. En *Proceedings of the 9th Mexican international conference on Advances in artificial intelligence: Part I*, 13-23.

da Cunha, I., J. Torres-Moreno y G. Sierra, 2011. On the Development of the RST Spanish Treebank. En *Proceedings of the 5th Linguistic Annotation Workshop*, 1-10.

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

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

Iruskieta, M., A. Díaz de Ilarraza y M. Lersundi, En prensa. Unidad discursiva y relaciones retóricas: un estudio acerca de las unidades de discurso en el etiquetado de un corpus en euskera. En *XXVII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*.

Karlsson, F., A. Voutilainen, J. Heikkila y A. Anttila, 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text.* Walter de Gruyter, .

Kiss, T. y J. Strunk, 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics,* 32 (4): 485-525.

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

Mann, W.C. y S.A. Thompson, 1987. Rhetorical Structure Theory: A Theory of Text Organization. *Text,* 8 (3): 243-281.

Marcu, D., 2000. *The theory and practice of discourse parsing and summarization.* The MIT press, Cambridge.

Pardo, T.A.S., 2006. SENTER: um segmentador sentencial automático para o português do Brasil. En: Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional:1-6.

Pardo, T.A.S. y M. Nunes, 2006. Review and Evaluation of DiZer–An Automatic Discourse Analyzer for Brazilian Portuguese. En *International Workshop on Computational Procesing of Written and Spoken Portuguese*, 180-189.

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

Pardo, T.A.S.y M.G.V. Nunes, 2002. Segmentação Textual Automática: Uma Revisão Bibliográfica. En: Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional.

Stede, M., 2004. The Potsdam commentary corpus. En *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, 96-102.

Taboada, M. J. Renkema, 2011. Discourse Relations Reference Corpus.

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

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

---

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

[ii] Más información sobre segmentadores en Pardo y Nunes (2002).

[iii] Utilizamos N-N (Núcleo-Núcleo) para señalar las relaciones paratácticas o relaciones multinucleares con más de un núcleo, mientras que utilizamos N-S (Núcleo-Satélite) para señalar las relaciones hipotácticas o relaciones nucleares con solo un núcleo, pudiendo ser su orden Núcleo-Satélite o Satélite-Núcleo.

[iv] La fuente de los ejemplos se indica primero por el acrónimo, seguido del año de publicación y un número que distingue los números publicados en un mismo año. Los artículos se han extraído de la página Web de la revista Gaceta Médica de Bilbao: http://www.elsevier.es/en/revistas/gaceta-medica-bilbao-316.

[v] El corpus anotado en diferentes niveles puede ser consultado en la página del grupo IXA dentro de la sección de recursos: https://ixa.si.ehu.es/Ixa/resources/Euskal_RSTTreebank.

[vi] En algunos casos algunos signos de puntuación (punto y dos puntos) pueden crear un segmento de discurso a pesar de no poseer un verbo conjugado.

[vii] MORPHEUS puede ser probado en: http://ixa2.si.ehu.es/demo/analisianali.jsp.

[viii] EUSTAGER puede probarse en: http://ixa2.si.ehu.es/demo/analisimorf.jsp.

[ix] EIHERA puede probarse en: http://ixa2.si.ehu.es/demo/entitateak.jsp.

[x] Utilizamos el carácter '[' para señalar el inicio de la segmentación manual y el ']' para el final. Y para la segmentación automática los caracteres '<' de inicio y '>' de final de segmento. Los caracteres de '{' y de '}' se utilizan para representar el inicio y final de la construcción SAME-UNIT en la segmentación manual.

[xi] Aunque esta regla es suficiente para la segmentación de A1, para la segmentación de A3 es necesario detectar la función sintáctica de *zerbitzu hauetako medikuek* 'el personal sanitario de estos servicios' que es el sujeto del verbo *jotzen dituzte* 'son consideradas' y, por tanto, parte del segmento.

[xii] Ofrecemos la primera acepción del diccionario OEH (http://www.euskaltzaindia.net/oeh): Ocuparse, estar en actividad; actuar, comportarse; hablar, tratar (sobre).

[xiii] En este caso la oración principal y la subordinada han sido detectadas y formalizadas correctamente tal y como se diseñaron para la segmentación automática, la cláusula adverbial subordinada dentro de la oración principal. Esa formalización es adecuada para las construcciones SAME-UNIT, cuando una EDU divide la otra EDU. Pero en este ejemplo no estamos ante tal construcción y pensamos que la formalización no coincide con la segmentación discursiva, ya que ambos segmentos se consideran EDUs y no hay un segmento que divida otro. Por lo tanto, de dicha formalización surgen dos diferencias: i) el primer segmento automático se considera sobre-segmentado, una composición de EDUs y ii) la marca de inicio del segundo segmento automático no coincide con el manual, se considera que hay una unidad que falta segmentar, en este caso una oración principal.

[xiv] La traducción de los ejemplos que se ofrece han sido extraídos del mismo artículo original, en los casos en los que la traducción se alejaba de la versión en euskera se ha modificado mínimamente acercándonos lo máximo posible a la explicación del fenómeno.