Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 41-48 recibido 26-03-2015 revisado 27-04-2015 aceptado 19-05-2015

EusEduSeg: A Dependency-Based EDU Segmentation for Basque

EusEduSeg: Un Segmentador Discursivo para el Euskera Basado en Dependencias

Mikel Iruskieta, Benat Zapirain
IXA Group. University of the Basque Country
{mikel.iruskieta, benat.zapirain}@ehu.eus

Resumen: Presentamos en este artículo el primer segmentador discursivo para el euskera (EusEduSeg) implementado con heurísticas basadas en dependencias sintácticas y reglas lingüísticas. Experimentos preliminares muestran resultados de más del 85 % F_1 en el etiquetado de EDUs sobre el Basque RST TreeBank.
Palabras clave: Segmentación discursiva, Rhetorical Structure Theory (RST), segmentador, euskera

Abstract: We present the first discursive segmenter for Basque implemented by heuristics based on syntactic dependencies and linguistic rules. Preliminary experiments show F_1 values of more than 85% in automatic EDU segmentation for Basque.
Keywords: Discourse segmentation, Rhetorical Structure Theory (RST), segmenter, Basque

1 *Introduction*

An obligatory first step in the annotation of any discourse parser is to identify the discourse units. This is known as the segmentation phase. The aim of segmentation is to mark the elementary units of the text, or in other words, to establish the basic elements of each language analysis level in order to enable the subsequent identification of the relation that exist between them.

The definition of an Elementary Discourse Unit (EDU) is nowadays controversial in the areas of Discourse Studies, and, as a consequence, several segmentation granularities (van der Vliet, 2010) have been proposed within RST[1].

Although it is hardly ever explicitly stated, segmentation proposals are based on the following three basic concepts:

- Linguistic "form" (or category).
- "Function" (the function of the syntactical components).
- "Meaning" (the coherence relation between propositions).

The possible combinations between these basic concepts used in discourse segmentation and those proposed in RST are underlined in Figure 1.

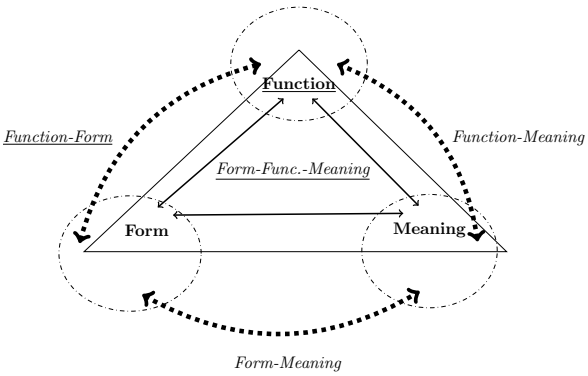

Figure 1: The basic concepts of discourse segmentation: form, function and meaning

Best-known segmentation proposals within RST are:

- The original RST proposal in English (Mann and Thompson, 1987): all clauses are EDUs, except for restrictive relative clauses and clausal subject or object components (syntactical function). This proposal is based solely on syntactical function.
- The first RST-based annotated corpus in English (Carlson and Marcu, 2001): in addition to that outlined in the orig-inal proposal, here both the compo-nents of attribution clauses (criterion based on function and meaning) and those phrases that begin with a discourse marker (e.g. *because of*, *spite of*, *accord-*

[1] A relational discourse structure theory proposed by Mann and Thompson (1987): for discourse coherence.

ISSN 1135-5948

ing to, etc.) are also segmented (criterion based on form and semantics).This proposal uses all three basic concepts: form, function and meaning.

– A segmentation proposal in English that adheres more closely to the original RST proposal (Tofiloski, Brooke, and Taboada, 2009): it segments verb clauses, coordinated clauses, ad-junct clauses and non-restrictive rela-tive clauses marked by a comma (it is a proposal based both on form restriction and syntactical function). Unlike in the proposal tabled by Carlson and Marcu (2001), in this method phrases beginning with discourse markers are not segmented, since they contain no verbs. In the annotation of the Spanish and Basque RST corpus, (da Cunha et al., 2010b; Iruskieta et al., 2013) this segmentation method was followed.

When attempting to define what a "discourse unit" actually is, these three basic concepts (form, function and meaning) pose a number of problems: *a*) If we based our analysis on form alone, many of the segmented elements would not be discourse units. *b*) If we based our analysis on function alone, then we would only be able to give annotators overly generalized definitions and imprecise segmentation criteria, such as adjunct clauses or adverbial clauses. *c*) And finally, if we based our analysis solely on meaning, we would encounter the problem of circularity between the segmentation annotation phase and the rhetorical relation annotation phase. The clearest example of this is that in order to annotate ATTRIBUTION relations, we would first have to segment the attribution clauses in the segmentation phase, resulting in a mixing of the two phases.

Following Thompson, Longacre, and Hwang (1985) we consider discourse units as functionally independent units, where three types of subordinate clauses can be distinguished: *i*) complements (which functions as noun phrases), *ii*) relative clauses (which functions as noun modifiers) and *iii*) adverbial clauses (which functions as modifiers of verb phrases or entire clauses). Blühdorn (2008) stated this subordinated but adverbial clauses can be seen as clause linkages, because it is the adverbial clauses which gives to the main clause a (discourse) thematic role.[2]

[2]More detailed information about adverbial clauses can be read in Liong (2000) and Lehmann (1985).

Clause type	**Example**
Independent sentence	[*Whipple (EW) gaixotasunak hesteei eragiten die bereziki.*]$_1$ GMB0503
Main, part of sentence	[pT1 tumoreko 13 kasuetan ez *zen* gongoila inbasiorik *hauteman*;]$_1$ [aldiz, pT1 101 tumoretatik 19 kasutan (18.6%) inbasioa *hauteman zen*, eta pT1c tumoreen artetik 93 kasutan (32.6%).]$_2$ GMB0703
Finite adjunct	[Haien sailkapena egiteko hormona hartzaileen eta c-erb-B2 onkogenearen gabeziaz baliatu gara,]$_1$ [*ikerketa anatomopatologikoetan erabili ohi diren zehaztapenak direlako.*]$_2$ GMB0702
Non-finite adjunct	[Ohiko tratamendu motek porrot eginez gero,]$_1$ [gizentasun erigarriaren kirurgia da epe luzera egin daitekeen tratamendu bakarra.]$_2$ GMB0502
Non-restrictive relative	[Dublin Hiriko Unibertsitateko atal bat da Fiontar,]$_1$ [zeinak Ekonomia, Informatika eta Enpresa-ikasketetako Lizentziatura ematen baitu, irlanderaren bidez.]$_2$ TERM23

Table 1: Main clause structures.

The segmentation guidelines we have use for Basque conflate all the approach presented before (Tofiloski, Brooke, and Taboada, 2009) and Basque clause combin-ing (Salaburu, 2012). As an example of what an EDU is, we show the main clause struc-tures in Table 1.

In this paper we present EusEduSeg[3] the first segmenter for Basque language, based on form and function rules. We evaluate the segmenter over a hand annotated corpora and we obtain promising results.

The remainder of this paper is structured as follows. Section 2 lays out the related work. Section 3 sets out the description of our system and Section 4 presents the experiment and results. Finally, Section 5 presents

[3]The segmenter EusEduSeg can be tested at http://ixa2.si.ehu.es/EusEduSeg/EusEduSeg.pl.

the discussion and establishes directions for future work.

2 Related Work

Although there are some works in Basque processing which identifies verbal chains, phrases (Aranzabe, 2008) and clauses (Aduriz et al., 2006), to cite some, there is not any discourse segmenter available for comparison in Basque. Iruskieta, Diaz de Ilarraza, and Lersundi (2011) established the bases for Basque discourse segmentation and implemented a prototypical segmenter reusing a statistical and morphological rule based chunk identifier (Arrieta, 2010). Including sentence boundaries, they obtained an F_1 of 66.94 in the experiments they car-ried out.

The evaluation of discourse segmentation it is not a trivial task, and several statistical measures have been used to check the robustness of a segmenter or to determine the reliability between human annotators and system evaluations:

i) Percent agreement was used to evaluate the agreement between human annota-tors by Hearst (1997) and Marcu (1999).

ii) Tofiloski, Brooke, and Taboada (2009) and Afantenos et al. (2010) used precision, recall and F_1 measures to evaluate the reliability and robustness of both automatic systems and human annotators.

iii) *Kappa* (κ) was used in Hearst (1997), Miltsakaki et al. (2004) and Tofiloski, Brooke, and Taboada (2009) to evaluate both automatic systems and human annotators.

Regarding to automatic discourse segmenters in languages others than Basque, Afantenos et al. (2010) presented a dis-course segmenter for French, da Cunha et al.(2010b) for Spanish and Tofiloski, Brooke, and Taboada (2009), Subba and Eugenio (2007) and Soricut and Marcu (2003) for En-glish. Table 2 summarizes the F_1 results pub-lished in those works.

Language	F_1	Reference
English	79	(Tofiloski, Brooke, and Taboada, 2009)
English	83-84	(Soricut and Marcu, 2003)
Spanish	80	(da Cunha et al., 2010a)
French	73	(Afantenos et al., 2010)

Table 2: State of the art in EDU parsing

The approach we followed to build our EDU segmentation system is rule-based and we avoid "same-unit" constructions as in Tofiloski, Brooke, and Taboada (2009). Specifically, as our rules are based on syntactical (dependencies) and morphological information, we follow a form-function approach for building our rule based automatic EDU segmentation.

3 EusEduSeg: System Description

From the syntactic point of view, most EDUs in the Basque RST TreeBank corpus exhibit two characteristic patterns that could be described as follows:[4]

- **Pattern 1**: verb nodes (*ROOT*, *ADI* and *ADT*) in the sentence's dependence tree govern an EDU if any of their recursively projected nodes accomplishes all the following conditions:

  1-*a*) It is the furthest node to the right from the governing head node (not necessarily the furthest one in the tree structure, but in the sentence order).

  1-*b*) It is a punctuation mark.

- **Pattern 2**: If a connector node (examples of LOT node are *edo* 'or', *eta* 'and', or *baina* 'but') has two direct verbal children nodes, then the connector node (LOT) delimits the frontier between two EDUs.

Given the simplicity of these dependency patterns, we developed a straightforward classifier that search for nodes that fulfill the previous conditions and label them as ending EDUs (E-EDU).

In order to better explain the patterns mentioned above, dependency trees in fig-ures 2 and 3 are introduced next. The tree in Figure 2 is a tree fragment (i.e. not the whole sentence's tree) representing an EDU that matches the pattern named as 1 right before. In this case, the node governing an EDU is the top most node in the tree, which is labeled as an verb (ADI) by Maltixa (Diaz de Ilarraza, Gojenola, and Oronoz, 2005), a dependency parser for Basque[5] (*lokalizatu* 'to

[4] Table 6 in Appendix A shows the descriptions of the Basque glosses employed in the paper.

[5] Maltixa can be tested at http://ixa2.si.ehu.es/maltixa/index.jsp.

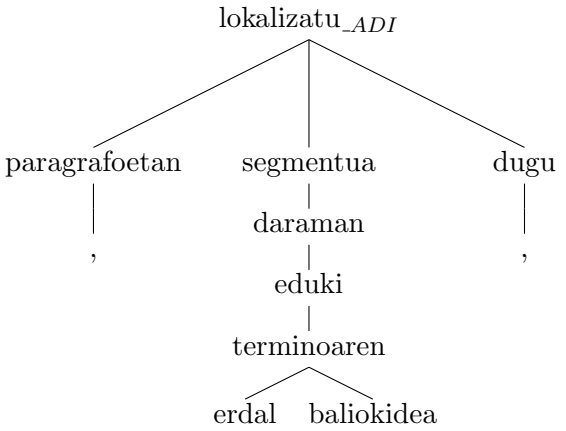

Figure 2: An application example of Pattern 1 (*paragrafoetan, erdal terminoaren eduki baliokidea daraman segmentua lokalizatu dugu,)* TERM28

locate'). As required by pattern 1, there is a punctuation node (a comma) under the auxiliary node *dugu* (auxiliar verb) that fulfills 1-a and 1-b conditions. This punctuation node is delimiting the frontier between the current EDU (represented in Figure 2) and the next one (the rest of the sentence is omitted here for lacking of space) and it should be labeled as an end-EDU (E-EDU) by the segmenter.

Figure 3 shows a tree fragment from the Basque RST TreeBank corpus that exactly matches pattern 2. There are two verbal nodes (ADI and ADT) and both share the same connector (LOT) parent node. As stated in pattern 2, the connector node es-tablishes boundaries between EDUs. In the example of Figure 3 the boundaries (E-EDU and B-EDU) would establish as follows (in bold): *...formal eta osoa lortzea lan neketsua* ***da***$_{E-EDU}$ ***eta***$_{B-EDU}$ *horretan datza atal...*

In order to increase the performance of the classifier, we added a post processing layer consisting of a rule set based on previous observations by Iruskieta, Diaz de Ilarraza, and Lersundi (2011). Target and token sequences that matches the target are underlined in corpus examples below:

- **Rule 1** (temporal): label *ADI (ERL:DENB)* nodes as *E-EDU*.

  (1) *Termino teknikoak hautatzerakoan*][*deklinabide kasua erabakigarria izan daiteke.* TERM31

- **Rule 2** (conditional-I): label *ERL:BALD + ,* sequences as *E-EDU*.

  (2) *Halako tresna bat euskararako garatu nahi badugu,*][*eragozpen gehiago topatuko dugu ondoko hiru arrazoiengatik.* TERM31

- **Rule 3** (conditional-II): tag *ERL:BALD + ere + ,* sequences as E-*EDU*.

  (3) *Emaitzarik ez badugu ere,*][*eredua izen-sintagmarena baino zabalagoa izango dela sumatzen dugu.* TERM31

- **Rule 4** (adjunct): label *ADI + ADB + ,* sequences as *E-EDU*.

  (4) *Ohiko tratamendu motek porrot eginez gero,*][*gizentasun erigarriaren kirurgia da epe luzera egin daitekeen tratamendu bakarra.* GMB0502

- **Rule 5** (reason): label *ERL:KAUS + ,* sequences as *E-EDU*.

  (5) *Hona hemen oin malgua izateagatik*][*kalkaneo-stop teknika erabiliz gure zerbitzuan ebakuntza egin diegun haurrek izandako emaitzak* GMB0601

- **Rule 6** (concessive): label *ERL:KONT* nodes as *E-EDU*.

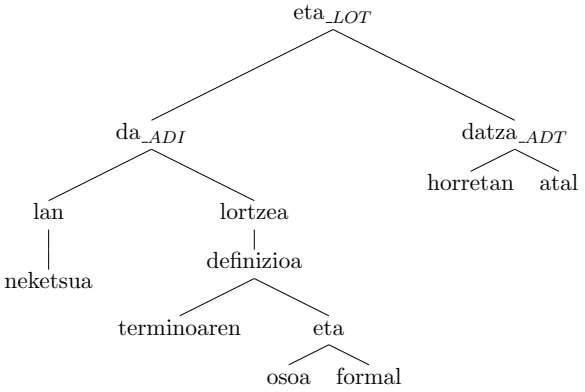

Figure 3: An application example of Pattern 2 (*Terminoaren definizio formal eta osoa lortzea lan neketsua eta horretan datza* [...] *atal* [...]) TERM31

(6) *Prebentzio metodoen eta artoplastiako teknika modernoen laguntzaz horrelako kasuak murriztu diren arren,*][*infekzio hori sendatzea erronka bat da oraindik ere.* GMB0802

- **Rule 7** (purpose): label *ADI(tzeko) + IZE + ,* as *E-EDU*.

(7) ingurunea aldatu ondoren elkarrekintza magnetikoak aztertzeko asmoz,][eta inguru biologikoetan ere erabiltzeko asmoz. ZTF17

4 Experiments and Results

4.1 Datasets

The corpus[6] used in this study consists of manually annotated abstracts from three specialized domains (medicine, terminology and science), and, it comprises 60 documents that contain 15,566 words (803 sentences) that were manually annotated with 1,355 EDUs and 1,292 relations. The corpus was analyzed with Maltixa, and randomly divided into training (50% for rule designing), development (25% for rule tuning) and test (25% for testing) sets.

4.2 EusEduSeg: EDU Segmenter

As mentioned before, the EDU classifier is entirely based on dependency and linguistic rules, as well as on a final consistency layer that checks the resulting EDUs with the aim of removing duplicated and incorrectly built EDUs (e.g: EDUs with no verbs in). In order to determine the influence of each rule set in the EDU segmentation task, we developed three different versions from the main classification system described in Section 3:

- ***EDU-Seg-1***: an EDU segmenter based only on dependency based patterns 1 and 2 described in Section 3.
- ***EDU-Seg-2***: an EDU segmenter based only on linguistic based rules (rules 1-7 from Section 3).
- ***EDU-Seg-3***: an EDU segmenter that takes advantage from both dependency based patterns and linguistic rules.

[6]The RST Basque Treebank (Iruskieta et al., 2013) and it's segmentation can be consulted at: http://ixa2.si.ehu.es/diskurtsoa/en/.

It is worth to remember that segmenter's rules and heuristics were developed manually and based, when needed, on observations made in training or development data.

EusEduSeg gives the possibility to configure several output formatting options that can be used in several tasks: *a*) web format to use in other NLP tasks. *b*) RSTTool format to annotate manually the RS-tree with RSTTool (O'Donnell, 2000). *c*) DiZer format (Pardo, Nunes, and Rino, 2004) to use in an automatic discourse parser.

System architecture is presented in Figure 4.

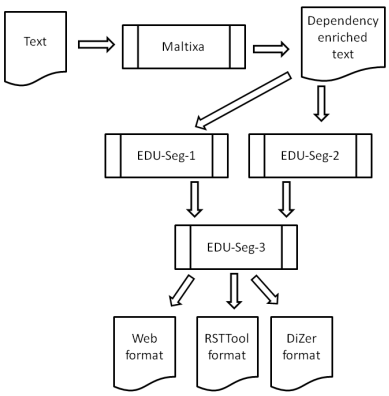

Figure 4: EusEduSeg phases

4.3 Evaluation measures

Performance of EDU segmenters has been reported with the standard precision, recall and F_1 measures, in similar way to many other authors on the task such as Tofiloski, Brooke, and Taboada (2009) and Afantenos et al. (2010). We calculate each of the mea-sures as follows:

$$precision = \frac{correct_{E-EDU}}{correct_{E-EDU} + excess_{E-EDU}}$$

$$recall = \frac{correct_{E-EDU}}{correct_{E-EDU} + missed_{E-EDU}}$$

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

where $correct_{E-EDU}$ is the number of correct *end-EDU*s, $excess_{E-EDU}$ is the number of overpredicted *end-EDU*s and $missed_{E-EDU}$ is the number of *end-EDU*s the system missed to tag.

Data set	correct	excess	missed	precision	recall	F_1	F_1'
Train	592	49	173	92.35	77.38	84.21	61.72
Dev	237	36	79	86.81	75.00	80.47	48.88
Test	292	25	95	92.11	75.45	82.95	60.52

Table 3: Results for EDU-Seg-1 on train, development and test sets

Data set	correct	excess	missed	precision	recall	F_1	F_1'
Train	548	14	217	97.5	71.63	82.59	53.89
Dev	208	9	108	95.85	65.82	78.04	30.76
Test	259	16	128	94.18	66.92	78.24	45.03

Table 4: Results for EDU-Seg-2 on train, development and test sets

Data set	correct	excess	missed	precision	recall	F_1	F_1'
Train	621	62	144	90.92	81.17	85.71	66.88
Dev	240	43	76	84.80	75.94	80.13	49.36
Test	303	39	84	88.59	78.29	83.12	62.61

Table 5: Results for EDU-Seg-3 on train, development and test sets

4.4 Results

Tables 3, 4 and 5 show the results ob-tained by EDU-Seg-1, EDU-Seg2 and EDU-Seg-3 respectively at the task of automatic segmentation of Basque texts. Correct, excess, missed, precision, recall and F_1 measures are reported, as customary for all data sets. The difference between F_1 and F_1' is that while former refers to classifier's F-score for all EDUs in the data set, latter refers to the F-score for "non trivial" EDUs only (hits on trivially identifiable EDU boundaries that begin or end a sentence are not take into account when computing F_1'). F_1' should be considered as the real indicator of the segmenter's performance.

Results show very high precision values for all segmenters used in the experiments. As already explained in previous sections, the heuristic and rule based engine of the segmenters makes this high precision values likely to be expected.

Regarding to the comparison between dependency based heuristics and linguistic rules (results shown in Table 3 and 4 respectively), linguistic rules are more precise than heuristics, but, on the other hand, higher recall values in Table 3 suggest that dependency based heuristics seem to be more general or better suited for broad spectrum EDU labeling.

Table 5 reports our best results in EDU segmentation experiments. The improvements in F_1 and F_1' with respect to the values in tables 3 and 4, seem to indicate that EDU-Seg-3 is able to successfully combine knowledge bases from EDU-Seg-1 and EDU-Seg-2, as well as that both dependency based heuristics and linguistic rules seem to be relatively complementary.

4.5 Error analysis

A more detailed error analysis, which is not under the scope of this work, will be useful for the future development of the automatic text segmentation of Basque text and also to improve Maltixa the automatic dependency analyzer for Basque.

A complex clause combining, as in Example 8, with three verbs (two coordinated finite verbs *erabakitzen dute* 'they decide it' and *jotzen dute* 'they go to' and one nominalized *jotzea* 'the going'), which can be detected with our system (Pattern 2), was not segmented by our system, due to some errors done by the dependency parser.

(8) Erabiltzaileen % 80ak bere kabuz erabakitzen dute larrialdi zerbitzu batetara jotzea][eta kontsulta hauen % 70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek.
*GMB*0401
The 80% of the users go by their own initiative to the emergency department,][and the 70% of the surgeries are

considered slights by the health staff. *TRANSLATION*

5 Conclusions and Future Work

In this paper we have introduced EusEduSeg, the first discourse segmenter for Basque implemented with simple dependency based heuristics and several high precision linguistic rules.

Experiments carried out on the Basque RST TreeBank corpus show competitive and promising results given the simplicity of the proposed solution and, in the same way, they leave enough room for improvement to more sophisticated and machine learning based architectures.

The authors are currently striving to achieve the following aims:

- To increase the performance of the segmenter adding more rules or better tuning the existing ones.
- To integrate a new layer of Constraint Grammar rules from previous work of Iruskieta, Diaz de Ilarraza, and Lersundi (2011).
- To train more sophisticated and robust classifiers by using state–of–the–art machine learning algorithms.
- To export the rule set of EusEduSeg into other languages such as English, Spanish or Portuguese. Given the lexical dependency of rules 1-7 from Section 3, this exportation task could be tough. However, patterns 1 and 2 seem more neutral and, thus, more suitable to be applied to other languages.

References

Aduriz, I., B. Arrieta, J.M. Arriola, A. Diaz de Ilarraza, E. Iza-girre, and A. Ondarra. 2006. Muga Gramatikaren optimizazioa (MuGa). Technical report, EHU.

Afantenos, S. D., P. Denis, P. Muller, and L. Danlos. 2010. Learning recursive segments for discourse parsing. In *Seventh conference on Inter-national Language Resources and Evalua-tion*, pages 3578–3584, Paris, France, 19-21 May.

Aranzabe, M. J. 2008. Dependentzia-ereduan oinarritutako baliabide sintak-tikoak: zuhaitz-bankua eta gramatika konputazionala. Doktore-tesia, Euskal Herriko Unibertsitatea, Donostia.

Arrieta, B. 2010. Azaleko sintaxi-aren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean. Doktore-tesia, Euskal Herriko Unibertsitatea, Donostia.

Blühdorn, H., 2008. *Subordination and coordination in syntax, semantics and discourse: Evidence from the study of connectives.* 'Subordination' versus 'Coordination' in Sentence and Text. Benjamins, Amsterdam.

Carlson, L. and Daniel M. 2001. Discourse tagging reference manual. Technical report.

da Cunha, I., E. SanJuan, J.M. Torres-Moreno, M. Lloberes, and I. Castellón. 2010a. Discourse segmentation for Spanish based on shallow parsing. In *9th Mexican international conference on Advances in artificial intelligence: Part I*, pages 13–23, Pachuca, Mexico, 8-13 November. Springer-Verlag.

da Cunha, I., E. SanJuan, J.M. Torres-Moreno, M. Lloberes, and I. Castellón. 2010b. Diseg: Un segmentador discursivo automatico para el español. *Procesamiento de Lenguaje Nat-ural*, 45.

Diaz de Ilarraza, A., K. Gojenola, and M. Oronoz. 2005. Design and Development of a System for the Detection of Agreement Errors in Basque. In *Computational Linguistics and Intelligent Text Processing*, pages 793–802. Springer.

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

Iruskieta, M., M. J. Aranzabe, A. Diaz de Ilarraza, I. Gonzalez, M. Lersundi, and O. Lopez de la Calle. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations. In *4th Workshop "RST and Discourse Studies"*, Brasil, October 21-23.

Iruskieta, M., A. Diaz de Ilarraza, and M. Lersundi. 2011. Bases para la implementación de un segmentador discursivo para el euskera. In *8th Brazilian Symposium in Information and Human Language Technology (STIL 2011),*

OCTOBER 2011.

Lehmann, C. 1985. Towards a typology of clause linkage. In *Conference on Clause Combining*, volume 1, pages 181–248.

Liong, T. 2000. Adverbial clauses, functional grammar, and the change from sentence grammar to discourse-text grammar. *Círculo de lingüística aplicada a la comunicación*, 4(2).

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

Marcu, D., 1999. *Discourse trees are good indicators of importance in text*, pages 123–136. Advances in Automatic Text Summarization. MIT, Cambridge.

Miltsakaki, E., R. Prasad, A. Joshi, and B. L. Webber. 2004. Annotating discourse connectives and their arguments. In *HLT/NAACL Workshop on Frontiers in Corpus Annotation*, pages 9–16, Boston, USA.

O'Donnell, M. 2000. RSTTool 2.4: a markup tool for Rhetorical Structure Theory. In *First International Conference on Natural Language Generation INLG '00*, volume 14, pages 253–256, Mitzpe Ramon, June12-16. ACL.

Pardo, T. A. S., M. G. V. Nunes, and L. H. M. Rino. 2004. DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese. *Advances in Artificial Intelligence–SBIA 2004*, pages 224–234.

Salaburu, P. 2012. Menderakuntza eta menderagailuak (Sareko Euskal Gramatika: SEG). http://www.ehu.es/seg/morf/5/2/2/2.

Soricut, R. and D. Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 149–156. Association for Computational Linguistics.

Subba, R. and B. Di Eugenio. 2007. Automatic discourse segmentation using neural networks. In *11th Workshop on the Semantics and Pragmatics of Dialogue*, page 189–190, Trento, Italy, 30-1 May-June.

Thompson, S. A., R. Longacre, and Shin Ja J. Hwang, 1985. *Adverbial clauses*, volume 2 of *Language Typology and Syntactic Description: Complex Constructions*, pages 171–234. Cambridge University Press, New York.

Tofiloski, M., J. Brooke, and M. Taboada. 2009. A syntactic and lexical-based discourse segmenter. In *47th Annual Meeting of the Association for Computational Linguistics*, pages 77–80, Suntec, Singapore, 2-7 August. ACL.

van der Vliet, N. 2010. Syntax-based discourse segmentation of Dutch text. In *15th Student Session, ESSLLI*, pages 203–210, Ljubljana, Slovenia, 1-12 August.

A Appendix: Glosses employed in the paper

Gloss abbrev.	Description
ADB	Adverb
ADI	Non-finite verb
ADL	Auxiliary finite verb
ADT	Finite verb
AUX	Auxiliary
BALD	Conditional clause
DENB	Temporal clause
ERL	Clause relation function
IZE	Noun
KAUS	Causal clause
LOT	Connector
PUNT	Punctuation
ROOT	Root of sentence

Table 6: Glosses used in examples