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 % F1 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 F1 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.

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

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 components 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-
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 segment-ation method was followed.

When attempting to define what a “dis-course 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 segmen-tation criteria, such as adjunct clauses or ad-verbial clauses. c) And finally, if we based our analysis solely on meaning, we would encoun-ter 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 an-notate 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 distin-guished: i) complements (which functions as noun phrases), ii) relative clauses (which functions as noun modifiers) and iii) adver-bial 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.\(^2\)

<table>
<thead>
<tr>
<th>Clause type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent sentence</td>
<td>[Whipple (EW) gaizotasunak hesteet eragiten die bereziki.](_1)</td>
</tr>
<tr>
<td>Main, part of sentence</td>
<td>[pT1 tumoreko 13 kasuetan ez zen gongoilua inbasiorik haute-manzan.](_1)</td>
</tr>
<tr>
<td></td>
<td>[aldir, pT1 101 tumore-tatik 19 kasutan (18.6%) inbasi-oa haute-manzan, eta pT1c tumoreen artetik 93 kasutan ((32.6%).](_2)</td>
</tr>
<tr>
<td>Finite adjunct</td>
<td>[Haien sailkapena egiteko hormona hartzaleen eta erb-B2 onkogenearen gabeziaz baliatu gara.](_1)</td>
</tr>
<tr>
<td></td>
<td>[ikerketa anato-mopatologikoetan erabili ohi diren zehaztapenak direlako.](_2)</td>
</tr>
<tr>
<td>Non-finite adjunct</td>
<td>[Ohiko tratamendu motek porrot eginez gero.](_1)</td>
</tr>
<tr>
<td></td>
<td>[gizentasun erigarriaren kirurgia da epe luzera egin daitekenez trata-mendu bakarra.](_2)</td>
</tr>
<tr>
<td>Non-restrictive relative</td>
<td>[Dublin Hiriko Unibertsita-tekateko atal bat da Fiontar.](_1)</td>
</tr>
<tr>
<td></td>
<td>[zeinak Ekonomia, Informatika eta Empresa-ikasketetako Lizentziatura ematen baitu, irlanderaren bidez.](_2)</td>
</tr>
</tbody>
</table>

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

---

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

\(^3\) The segmenter EusEduSeg can be tested at [http://ixa2.s1.ehu.es/EusEduSeg/EusEduSeg.pl](http://ixa2.s1.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. Iruskietx, Diaz de Ilarraza, and Lersundi (2011) established the bases for Basque discourse segmentation and implemented a prototypical segmenter reusing a statistical and morphologica rule based chunk identifier (Arrieta, 2010). Including sentence boundaries, they obtained an F1 of 66.94 in the experiments they carried 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 annotators by Hearst (1997) and Marcu (1999).

ii) Tofiloski, Brooke, and Taboada (2009) and Afantenos et al. (2010) used precision, recall and F1 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 English. Table 2 summarizes the F1 results pub-lished in those works.

<table>
<thead>
<tr>
<th>Language</th>
<th>F1</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>79</td>
<td>(Tofiloski, Brooke, and Taboada, 2009)</td>
</tr>
<tr>
<td>English</td>
<td>83-84</td>
<td>(Soricut and Marcu, 2003)</td>
</tr>
<tr>
<td>Spanish</td>
<td>80</td>
<td>(da Cunha et al., 2010a)</td>
</tr>
<tr>
<td>French</td>
<td>73</td>
<td>(Afantenos et al., 2010)</td>
</tr>
</tbody>
</table>

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 syntac-tical (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:

- **Pattern 1:** verb nodes (ROOT, ADI and ADT) in the sentence’s dependence tree govern an EDU if any of their recursive 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’, etxa ‘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 figures 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 Maltix (Diaz de Ilarraza, Gojenola, and Oronoz, 2005), a dependency parser for Basque5 (lokalizatu ‘to’)

---

3Table 6 in Appendix A shows the descriptions of the Basque glosses employed in the paper.

5Maltix can be tested at [http://ixa2.si.ehu.es/maltix/index.jsp](http://ixa2.si.ehu.es/maltix/index.jsp).
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 establishes boundaries between EDUs. In the example of Figure 3 the boundaries (E-EDU and B-EDU) would establish as follows (in bold): ...

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 hautatzerakoak [dekinabide kasua era] erabaki
garria izan daiteke. TERM31

- **Rule 2** (conditional-I): label ADI + ADB + , sequences as E-EDU.
  2. Halako tresna bat euskararako garatua nahi dugu, | eragozpen gehiago topatuko dugu ondoko hiru arrazoiengatik. TERM31

- **Rule 3** (conditional-II): tag ERL:BALD + ere + , sequences as E-EDU.
  3. Emaitzariak ez badugu ere, | eredu izen-sintagmarenai baino zabalagoa izango dela sumatzen dugu. TERM31

- **Rule 4** (adjunct): label ADI + ADB + , sequences as E-EDU.
  4. Ohiko tratamendu motek porrot eginez gero, | giztasun erigarriaren kirurgia da epe luzera egin daitekeen tratamendu bakarra. GMB0502

- **Rule 5** (reason): label ERL:KAUS + , sequences as E-EDU.
  5. Hona hemen oin maligua izateagatik | kalkaneo-stop teknika erabiltsez gure zerbitzuak ebakuntza egin diegun haurrek izandako emaitzak GMB0601

- **Rule 6** (concessive): label ERL:KONT nodes as E-EDU.
(6) Prebentzio metodoen eta artoplastiako teknika modernoan laguntzaz horrelako kasuak murriztu diren arren, || infekzio hori sendatzea erronka bat da oraindik ere. GMB0802

- **Rule 7** (purpose): label 
  \[\text{ADI(tzeko)} + \text{IZE} + , \text{as E-EDU}.\]

(7) ingurunea aldatu ondoren elkarrekintza magnetikoak aztertzeo asmoz, || eta inguru biologikoa 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.

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](image)

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 measures as follows:

\[
\text{precision} = \frac{\text{correct}_{E-EDU}}{\text{correct}_{E-EDU} + \text{excess}_{E-EDU}}
\]

\[
\text{recall} = \frac{\text{correct}_{E-EDU}}{\text{correct}_{E-EDU} + \text{missed}_{E-EDU}}
\]

\[
F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

where \text{correct}_{E-EDU} is the number of correct \text{end-EDUs}, \text{excess}_{E-EDU} is the number of overpredicted \text{end-EDUs} and \text{missed}_{E-EDU} is the number of \text{end-EDUs} the system missed to tag.

\(^6\)The RST Basque Treebank (Iruskia et al., 2013) and it’s segmentation can be consulted at: [http://ixa2.si.ehu.es/diskurtsoa/en/](http://ixa2.si.ehu.es/diskurtsoa/en/).
Table 3: Results for EDU-Seg-1 on train, development and test sets

<table>
<thead>
<tr>
<th>Data set</th>
<th>correct</th>
<th>excess</th>
<th>missed</th>
<th>precision</th>
<th>recall</th>
<th>F₁</th>
<th>F₁'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>592</td>
<td>49</td>
<td>173</td>
<td>92.35</td>
<td>77.38</td>
<td>84.21</td>
<td>61.72</td>
</tr>
<tr>
<td>Dev</td>
<td>237</td>
<td>36</td>
<td>79</td>
<td>86.81</td>
<td>75.00</td>
<td>80.47</td>
<td>48.88</td>
</tr>
<tr>
<td>Test</td>
<td>292</td>
<td>25</td>
<td>95</td>
<td>92.11</td>
<td>75.45</td>
<td>82.95</td>
<td>60.52</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Data set</th>
<th>correct</th>
<th>excess</th>
<th>missed</th>
<th>precision</th>
<th>recall</th>
<th>F₁</th>
<th>F₁'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>548</td>
<td>14</td>
<td>217</td>
<td>97.5</td>
<td>71.63</td>
<td>82.59</td>
<td>53.89</td>
</tr>
<tr>
<td>Dev</td>
<td>208</td>
<td>9</td>
<td>108</td>
<td>95.85</td>
<td>65.82</td>
<td>78.04</td>
<td>30.76</td>
</tr>
<tr>
<td>Test</td>
<td>259</td>
<td>16</td>
<td>128</td>
<td>94.18</td>
<td>66.92</td>
<td>78.24</td>
<td>45.03</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Data set</th>
<th>correct</th>
<th>excess</th>
<th>missed</th>
<th>precision</th>
<th>recall</th>
<th>F₁</th>
<th>F₁'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>621</td>
<td>62</td>
<td>144</td>
<td>90.92</td>
<td>81.17</td>
<td>85.71</td>
<td>66.88</td>
</tr>
<tr>
<td>Dev</td>
<td>240</td>
<td>43</td>
<td>76</td>
<td>84.80</td>
<td>75.94</td>
<td>80.13</td>
<td>49.36</td>
</tr>
<tr>
<td>Test</td>
<td>303</td>
<td>39</td>
<td>84</td>
<td>88.59</td>
<td>78.29</td>
<td>83.12</td>
<td>62.61</td>
</tr>
</tbody>
</table>

### 4.4 Results

Tables 3, 4 and 5 show the results obtained 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 F₁ measures are reported, as customary for all data sets. The difference between F₁ and F₁’ 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 F₁’). F₁’ 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 F₁ and F₁’ 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 Maltixia the automatic dependency analyzer for Basque.

A complex clause combining, as in Example 8, with three verbs (two coordinated finite verbs erabakitsen dute ‘they decide it’ and jotzen dute ‘they go to’ and one nominalized jotzena ‘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 erabakitsen dute larrialdi zerbitzu bate-tara jotzena || eta kontsulta hauen % 70a larritasun gutxikotzat jotzen dituzte zerbitzu hauetako medikuek.  

**GMB80401**  
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
OCTOBER 2011.


A Appendix: Glosses employed in the paper

<table>
<thead>
<tr>
<th>Gloss abbrev.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADB</td>
<td>Adverb</td>
</tr>
<tr>
<td>ADI</td>
<td>Non-finite verb</td>
</tr>
<tr>
<td>ADL</td>
<td>Auxiliary finite verb</td>
</tr>
<tr>
<td>ADT</td>
<td>Finite verb</td>
</tr>
<tr>
<td>AUX</td>
<td>Auxiliary</td>
</tr>
<tr>
<td>BALD</td>
<td>Conditional clause</td>
</tr>
<tr>
<td>DENB</td>
<td>Temporal clause</td>
</tr>
<tr>
<td>ERL</td>
<td>Clause relation function</td>
</tr>
<tr>
<td>IZE</td>
<td>Noun</td>
</tr>
<tr>
<td>KÄUS</td>
<td>Causal clause</td>
</tr>
<tr>
<td>LOT</td>
<td>Connector</td>
</tr>
<tr>
<td>PUNT</td>
<td>Punctuation</td>
</tr>
<tr>
<td>ROOT</td>
<td>Root of sentence</td>
</tr>
</tbody>
</table>

Table 6: Glosses used in examples