Detection of Central Units in Basque Scientific Abstracts*

La detección de la unidad central de resúmenes cientificos en euskera

Mikel Iruskieta, Arantza Díaz de Ilarraza, Gorka Labaka y **Mikel Lersundi**
IXA Group. University of the Basque Country
mikel.iruskieta@ehu.eus, a.diazdeilarraza@ehu.eus, gorka.labaka@ehu.eus, mikel.lersundi@ehu.eus

Resumen: En este artículo presentamos un detector automático basado en reglas que detecta la unidad discursiva más importante de un resumen cientifico. La detección de la unidad central es, tras la segmentación, un estadío de anotación crucial de la *Rhetorical Structure Theory* (RST) que puede ser explotado tanto en tareas de resumen automático como en tareas de búsqueda de preguntas. Los resultados demuestran que las unidades centrales de resúmenes científicos en euskera pueden ser detectadas automáticamente, aunque todavía hay espacio para mejora.
Palabras clave: Tópico discursivo, unidad central, RST

Abstract: This paper presents an automatic rule-based detector of the most salient discourse units in scientific abstracts. After segmentation, the detection of the central unit is a crucial annotation phase in the *Rhetorical Structure Theory* (RST), which could be exploited in automatic summarization or question answering tasks. Although there is still room for improvement, our results show that the central unit can be detected in Basque scientific abstracts.
Keywords: Discourse topic, central unit, RST

1 Introduction

Language users know how to identify the global meaning of the text (van Dijk, 1980). Detecting the global meaning and its relations with local meaning is very important to develop advanced NLP applications such as question answering, automatic summarization and sentiment analysis.

The global meaning is a kind of summary of the text which can have different forms: keyword (a single word); title (a phrase without a main verb); discourse topic, thematical sentences (van Dijk, 1980) or thesis statement (Burstein et al., 2001) (a sentence);[1] central proposition (Pardo, Rino, and Nunes, 2003) (a proposition extracted from the text) and central subconstituent (Egg and Redeker, 2010) or central unit (Stede, 2008) (in RST the most salient node of the RS-Tree).

The aim of this paper is to establish the basis for the automatic detection of the central unit. Before explaining our proposal, let us show the differences between the thesis statement and the central unit.

Burstein et al. (2001) have defined a thesis statement as follows:

> A thesis statement is defined as the sentence that explicitly identifies the purpose of the paper or previews its main ideas. (...) thesis statements reflect the most important sentences in essays.
>
> (Burstein et al., 2001, 99-100)

According to Paice (1980), most of the time the thesis statement is pointed to the reader by some indicators.

The Central Unit (CU) is a concept associated to the RS-trees that can be defined as an Elementary Discourse Unit (EDU) that has as special function to be the main nucleus in the tree. Elements attached to the CU are attached as satellites, which never acts as satellite in any relation. In contrast to Burstein et al. (2001), in an RS-tree there will always be (at least) an EDU that works as a central unit (even in cases where the thesis statement is elided).[2]

* This study was carried out within the framework of the following projects: Ber2Tek (IE12-333); NewsReader project (FP7- ICT-2011-8- 316404); IXA group, Research Group of type A (IT344-10).

[1] Sometimes the discourse topic has to be created by the reader, because it is implicit.

[2] There is no difference between *thesis statement* and *central unit* in Example (1). A clear difference

Example (1) shows a tagged text of the medical domain extracted from our corpus.

(1) [**Estomatitis Aftosa Recurrente (I): Epidemiologia, etiopatogenia eta aspektu klinikopatologikoak.**]$_1$ ["Estomatitis aftosa recurrente" deritzon patologia, ahoan agertzen den ugarienetako bat da,]$_2$ [tamainu, kokapena eta iraunkortasuna aldakorra izanik.]$_3$ [Honen etiologia eztabaidagarria da.]$_4$ [Ultzera mingarri batzu bezala agertzen da,]$_5$ [hauek periodiki beragertzen dira.]$_6$ [Lan honetan patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko garrantsitsuenak analizatzen ditugu.]$_7$ **GMB0301**[3]

Example (1) was annotated by two annotators. It was segmented in 7 EDUs and both annotators (A_1 and A_2) identified the last EDU (EDU_7) as the main EDU (the CU).

Paice (1980) states that 28 out of the 32 abstracts they studied have the thesis statement. Burstein et al. (2001) also detected that 7% of the texts do not have an explicit thesis statement. Paice (1980) categorizes the thesis statement indicators as follows: nouns (*paper*, *article*, *presentation*, *investigation*, *method*, *result*...), verbs (*discuss*, *introduce*, *present*, *examine*, *analy-*, *stud-*...), demonstratives (*this*, *the*, *a*, *some*...) and some pronouns (*we*, *I*...).

Following Paice (1980) in Example (1) we identify the following indicators: *i*) *Lan honetan* 'in this work' in Basque, the demonstrative *hau* 'his' refers to the work the writers are presenting. *ii*) The adjective *garrantzitsu* 'important' and the superlative *-en-* 'the most' indicate that this sentence is prominent in the text. *iii*) The verb *analizatu* 'analyze' is a common verb for expressing the main action of a piece of research (Iruskieta, de Ilarraza, and Lersundi, 2014). Its meaning is associated with the wordnet synset 'analyze$_1$'.[4] *iv*) The pronoun adjoined to the auxiliary of the verb, *-gu* 'we', shows that the topic the writers are referring to is an action performed by themselves. These indicators will give us some cues to identify the central unit automatically, even if they might be ambiguous.

The final aim of this work is to build a detector of the central unit to be used in different NLP applications for the Basque language. To detect the central unit of a text automatically we have used the information contained in the Basque RST Treebank.[5] We present a rule based detector of the central unit and the results obtained in the corpus.

The remainder of this paper is structured as follows. Section 2 lays out the related work and the theoretical framework and Section 3 the methodology used to build the detector of the central unit. Section 4 presents the system and Section 5 sets out the results of the detector. Finally, in Section 6 we present the discussion and directions for future work.

2 *Related work*

For the task of extracting the most relevant unit, Neto et al. (2000) use a text mining method, while Luhn (1958) is based on keywords. Pardo, Rino, and Nunes (2003) extract the gist sentence based on keywords and on text mining in Portuguese and English scientific text, where the former method significantly outperforms the latter.

In Burstein et al. (2001) two professional writers annotated the thesis statement of 100 texts and the agreement between both was 71% F-score and 0,733 κ. After that they reached almost the same results with machine learning techniques.

Our work is similar to Luhn (1958) and Burstein et al. (2001). To our best knowledge, this is the first proposal of a central unit detector for Basque texts. In Iruskieta, de Ilarraza, and Lersundi (2014) we find some considerations about the annotation of CUs as part of a general discourse annotation strategy. They observed that if CUs are

between both terms can be found in Example (3).

[3]Original translation of the Example (1): [**Recurrent aphthous stomatitis (I): epidemiologic, ethiologic and clinical features.**]$_1$ [Recurrent aphtous stomatitis is one of the most frequent oral pathology.]$_2$ [It has a controversial ethiology]$_4$ [and it is characterized by the apparition of painful]$_5$ [and recurrent ulcers with a variable size, location and duration.]$_3$ [In this paper we analyze the most important epidemiological, ethiological, pathological and clinical features of this common oral pathology.]$_7$

[4]It belongs to the reasoning category determined by the SUMO ontology.

[5]The RST Basque Treebank (Iruskieta et al., 2013) can be consulted at `http://ixa2.si.ehu.eus/diskurtsoa/`.

previously annotated, the degree of annotator agreement in the RS-trees is greater.

3 Methodology

The corpus used in this paper (see Table 1 consists of abstracts from five specialized domains (medicine, terminology, science, health and life) collected by UZEI[6] and the Summer Basque University (UEU)[7] as organizers of conferences in those areas.

Corpus	Domain	Source
GMB	Medicine	Gaceta Médica de Bilbao (2000 - 2008)
TERM	Terminology	Int. Conference on Terminology, 1997 organized by UZEI
ZTF	Science	Scientific articles Faculty of Science UPV/EHU
OSA	Health	2nd Symposium of Basque Researches, 2014, UEU
BIZ	Life	1st Symposium of Basque Researches, 2010, UEU

Table 1: Corpus description: Domains and Sources

The gold standard we created contains 25,593 EDUs and 100 texts, each with its CUs. A more detailed description is presented in Table 2.

Corpus	Texts	Words	EDUs
GMB	20	3,010	283
TERM	20	5,664	584
ZTF	20	6,892	603
OSA	20	4,878	475
BIZ	20	5,535	569
Total	**100**	**25,593**	**2,514**

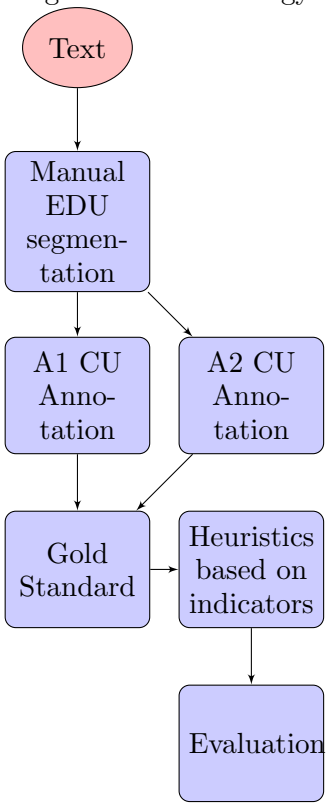
Table 2: Corpus description: measures

The corpus we have used is bigger or similar to others created for similar aims. Paice (1980) used a corpus of 32 texts and Burstein et al. (2001) used a corpus of 100 texts. We have used the GMB, TERM and ZTF subcorpora as a training data-set and the OSA and BIZ corpora as test data-sets.

The corpus was annotated by two linguists who were familiar with the RSTTool.[8] The annotation phases represented in Figure 1 were as follows:

i) Annotators segmented the texts manually following Iruskieta, Diaz de Ilarraza, and Lersundi (2011).
ii) Both annotators determined the CU of each text.
iii) The results were evaluated and harmonized following Iruskieta (2014).
iv) Some indicators were manually extracted (Iruskieta, 2014).
v) Heuristics that exploit these indicators were defined.
vi) The results were evaluated.

Figure 1: Methodology



3.1 Agreement between annotators

The inter-annotator agreement (A_1 and A_2) are comparable to Burstein et al. (2001)[9] with a Kappa score of 0.796 (for a total of 2440 EDUs).

The most common disagreements between annotators were the following:

– EDUs annotated as CUs: annotators

[6] http://www.uzei.eus/.
[7] http://www.ueu.eus/.
[8] http://www.isi.edu/licensed-sw/RSTTool/.
[9] The agreement in Burstein et al. (2001) was kappa 0.733 (for a total of 2,391 sentences) in 100 texts between two annotators.

judged differently the importance of the EDU(s) to be considered as CU(s).

In Example (2) the central unit consists of EDU_1 (where the paper proposes new terminology) and EDU_3 (where the paper reports on aspects of the tool). This was confirmed also in the following annotation phase, in the labeling of the relation, because both EDUs were linked using the CONTRAST multinuclear relation. Otherwise, EDU_2 could not be part of the central unit, because it was labeled as a satellite of the first EDU.

Example (2) shows the case in which annotator A_1 and A_2 identified different EDUs as CU (for A_1 the first EDU is the CU while for A_2 the CU is the first, the second and the third EDU).

(2) [Artikulu honetan, terminologia eleanitza sortzeko metodologia bat proposatuko dugu,]$_1$ *A*1&*A*2 [orain arte izan ditugun esperientzietan oinarrituta;]$_2$ *A*2 [baina tresnaren beste alderdi batzuk ere azalduko ditugu.]$_3$ *A*2 [...] **TERM39**[10]

– When the topic is not explicit, the annotation of CUs differs severely as shown in Example (3) where *A*1 and *A*2 annotated different EDUs (as can be seen highlighted in the text).

(3) **Energiarako materialak: Litio-ioi bateriak.**
Litio-Ioi bateriak ezinbestekoak dira gure eguneroko bizitzan. Ez al duzu telefono mugikorrik, mp3rik ala ordenagailu eramangarririk zeure poltsan? Bateria hauen arazorik handienetarikoak pisua eta bolumena dira. Gainera, bere osagaiak prozesatu behar dira bere toxikotasunarengatik.
[Ikerketa-ildo hau, solido egoeraren kimika aztertzen duen taldean sortu da.]*A*2 [*Bere helburua LiFePO*$_4$ *materialaren optimizazioa da*,]*A*1 litio-ioi bateria komertzialetan konposatu hau katodo moduan erabiltzeko. Gaur egun erabiltzen den katodoarekin, $LiCoO_2$-rekin konparatuta ($LiCoO_2$), aukeratu den konposatua ez da kutsagarria, energi densitate handia dauka, seguruagoa da, eta pisu baxuagoa eta prezio hobea ditu. [...] **ZTF18**[11]

In case of disagreement a superannotator harmonized the annotations establishing a general criteria for the CU:

i) The thematic sentence, when it is explicit.

ii) When the thematic sentence is not explicit, the harmonization criteria for establishing the CU establishes the CU based on the criteria that subjects are possible candidates: the aim of the research, method, results and conclusions (in this order).

In Example (3) the annotation proposed by *A*2 was excluded because it modifies the central unit.

After the creation of the gold standard corpus, we looked for elements that were candidates to be indicators of the Central Unit. We concentrated on verbs, nouns,[12] pronouns and bonus words.

We enlarged the list of indicators proposed by Paice (1980). The resulting list is shown in Table 3. We have highlighted in gray those indicators that were not accounted for in the mentioned work.

[10] Translation: [This paper will propose a methodology for sourcing multilingual terminology]$_1$ [based upon our experiences to date,]$_2$ [but also report on other aspects of the tool.]$_3$ [...].

[11] Translation: **Energy materials: Lithium–ion batteries.**
In the last few years, lithium–ion batteries have become essential in daily life, don't you have a mobile, a mp3 player or even a laptop in your bag? Examples of their use are in fact countless. Yet, and despite their widespread use, these batteries present two major disadvantages: they are heavy and bulky, and require an expensive recycling process at the end of their lifecycle because of the toxicity of some of their components, such as Co. [The present research line was born in a solid–state devoted research group,]*A*2 [with the aim of materializing the use of LiFePO 4]*A*1 as commercial cathode in lithium–ion batteries. The choice of this material to replace currently used LiCoO 2 is supported by its lack of toxicity, low contaminant potential, high energy density, lower price and more secure operation. [...].

[12] Verbs and nouns have their corresponding synset of the Basque Wordnet associated. http://adimen.si.ehu.eus/cgi-bin/wei/public/wei.consult.perl.

Verbs		Nouns	
BSQ	**ENG$_{MCR}$**	**BSQ**	**ENG$_{MCR}$**
aztertu	examine$_1$	abiapuntu$_1$	starting_point$_1$
analizatu	examine$_1$	arlo$_1$	subject_field$_1$
oinarritu	base$_1$	artikulu$_7$	article$_1$
baloratu	value$_2$	asmo$_2$	purpose$_1$
azaldu	recount$_1$	bide$_2$	means$_1$
aurkeztu	present$_2$	gai$_6$	topic$_1$
aipatu	present$_2$	ikerkuntza$_3$	
berri eman	present$_2$	ikerketa$_2$	research$_2$
jardun	present$_2$	azterlan$_3$	
plazaratu	present$_2$	ikerlan$_3$	
ikertu	investigate$_1$	arazo$_3$	problem$_2$
erabili	use$_1$	irtenbide$_2$	resolution$_4$
Demonstrative Pronouns		komunikazio	paper$_5$
hau	this	hitzaldi$_2$	speech$_1$
Personal Pronouns		lan$_3$	work$_2$
gu	we	lan-ildo	−−
Bonus Words		lerro$_{11}$	line$_8$
garrantzi(tsu)	important	ikerketa-lerro	
nagusi	main	proiektu$_2$	project$_2$
azpimarragarri	remarcable	ikerketa-proiektu	
eskerga	huge	talde$_1$	group$_1$
(gaur) egun	nowadays	ikerketa-talde	
		xede$_1$	goal$_1$
		helburu$_2$	

Table 3: Indicators extracted from the central units with WordNet synsets

These indicators helped us define the heuristics to be implemented in our automatic CU detector system.

4 *The system*

Based on the indicators of Table 3 we implemented eight different heuristics (two of them are combinations of the others).

Each heuristic is applied under the limits of an EDU. Before showing the results, let us first explain the defined heuristics.

- Heuristic-1 (nouns and verbs) considers as central unit those EDUs which contain any of the nouns or verbs marked as indicators in our empirical study or combinations of both.
- Heuristic-2 (nouns and verbs + pronouns) identifies as central unit those EDUs with a combination of nouns plus demonstrative pronouns (e.g. the demonstrative pronoun *hau* 'this') within a three-word distance or verbs accompanied with a first plural personal pronoun or in the first person plural.[13]
- Heuristic-3 (nouns and verbs + bonus words) considers as central unit those EDUs with a combination of nouns, verbs and bonus words.
- Heuristic-4 (nouns and verbs + title words) considers as central unit those EDUs with a combination of two nouns, one verb and any other word from the title of the document.
- Heuristic-5 (position in the document) considers central units those EDUs which are at the beginning or in the middle of the text. In the case of long texts, the last EDU considered is the EDU in position 20.
- Heuristic-6 (main verb) considers as central unit those EDUs with a main finite verb.
- Heuristic-7 combines heuristics 1, 2 and 4 to identify as central unit those EDUs that satisfies any of this constraints: *i*) a noun, *ii*) a noun with a demonstrative pronoun which is within a three words distance, and *iii*) a title word with a noun or a verb.
- Heuristic-8 combines heuristics 1, 2, 3, 4 and 5. So it considers only the EDUs from 1 to 20 identifies as central unit those EDUs that satisfies any of this constraints: *i*) a noun with a demonstrative pronoun which is within a three words distance, *ii*) a bonus word with a noun or a verb, *iii*) a word from the title with two nouns and a verb, and *iv*) a verb with a personal pronoun.

[13] In Basque, we analyzed two types of pronouns: *a*) The first person plural pronoun *gu* 'we' and *b*) the first person plural embedded in auxiliary verbs.

5 *Results*

The performance of the heuristics is reported following the standard measures precision, recall and f-score (F_1). We calculate each of the measures as follows:

$$precision = \frac{correct_{CU}}{correct_{CU} + excess_{CU}}$$

$$recall = \frac{correct_{CU}}{correct_{CU} + missed_{CU}}$$

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

where $correct_{CU}$ is the number of correct *central units* (C), $excess_{CU}$ is the number of overpredicted *central units* (E) and $missed_{CU}$ is the number of *central units* the system missed (M).

Table 4 shows the results obtained on the training and development sets. As we reported in Table 2, out of a total of 1,210 EDUs there are 79 central units on the training and development sets.

Data set	C	E	M	Prec.	Rec.	F_1
Heuristic-1	19	44	60	0.30	0.24	0.27
Heuristic-2	36	63	43	0.36	0.46	0.40
Heuristic-3	20	37	59	0.35	0.25	0.29
Heuristic-4	10	4	69	**0.71**	0.13	0.22
Heuristic-5	24	33	55	0.42	0.30	0.35
Heuristic-6	69	840	10	0.08	**0.87**	0.14
Heuristic-7	34	43	45	0.44	0.43	0.44
Heuristic-8	50	65	29	0.43	0.63	**0.52**

Table 4: Results for all the heuristics on the training and development sets

We can observe in Table 4 that we have improved the results in both combinations: Heuristic-7 and Heuristic-8. In Heuristic-8 we have combined almost all the heuristics (except Heuristic-6), obtaining the best f-measure. There are three heuristics that have a better partial result:

i) Heuristic-4 and Heuristic-7 are better in precision, which means that those heuristics are more precise, that, is they label more correct CUs than excess ones. The combination of a title word and a 'noun + verb' structure mainly appear in the CU. The combined Heuristic-7 is slightly better in precision, but it misses many more CUs than Heuristic-8.

ii) Heuristic-6 is better in recall, which means that this heuristic missed less CUs than Heuristic-8, but it labels more wrong CUs.

Once we tested different combinations, we chose those with the better results: Heuristic-7 and Heuristic-8. Table 5 shows the results obtained on the test set consisting of 44 central units.

Data set	C	E	M	Prec.	Rec.	F_1
Heuristic-1	15	31	29	0.33	0.34	0.33
Heuristic-2	22	68	22	0.24	0.50	0.33
Heuristic-3	5	14	39	0.26	0.11	0.16
Heuristic-4	7	3	37	**0.70**	0.16	0.26
Heuristic-5	40	711	4	0.05	0.91	0.10
Heuristic-6	41	721	3	0.05	**0.93**	0.10
Heuristic-7	21	30	23	0.41	0.48	**0.44**
Heuristic-8	23	48	21	0.32	0.52	0.40

Table 5: Results for all the heuristics on test sets

Table 5 shows the results obtained by all the heuristics for the detection of the central unit of Basque scientific abstracts. Although we are far from the results obtained by human annotators —the agreement between two annotators in the test set was a f-score of 0.89—,[14] the results of Table 5 obtained on test set show that Heuristic-8 was overfitted to the training corpus and, in contrast, Heuristic-7 maintains the same results in the test corpus, even in different domains.

In order to see how our system works, we have compared the performance of the central unit detection only over the cases that the annotators agreed, since we know that a number of cases (7 texts) are ambiguous also for the human annotators. The results are presented in Table 6.

Data set	C	E	M	Prec.	Rec.	F_1
Heuristic-7	19	19	16	0.50	0.54	**0.52**
Heuristic-8	20	35	15	0.36	0.57	0.44

Table 6: Results for combined heuristics, excluding ambiguous texts for human annotators

As we can see in Table 6 the combinations done in Heuristic-7 seem to indicate that this heuristic is able to successfully combine the Heuristic-1, Heuristic-2 and the Heuristic-4 to improve results.

6 *Conclusions and future work*

This paper presents the first central unit detector for Basque. The heuristics implemented in the system have been defined based on the analysis performed on the RST Basque TreeBank. Although we reach similar promising results for different domains, we believe that there is still room to improve the heuristics and the combination strategies. Furthermore, the gold standard files used to evaluate the system are available for anyone to use at the RST Basque TreeBank.[15]

The work carried out will be useful for adding discourse hierarchy information in certain language processing tasks for Basque, as such as question answering (Aldabe, 2011), automatic summarizers and sentiment analysis (Alkorta et al., 2015).

[14] And a 0.95 of κ.

[15] The files can be download from http://ixa2.si.ehu.eus/diskurtsoa/fitxategiak.php.

6.1 Future work

The authors are currently developing a system that works with machine learning techniques and aim to study the possibility of combining both systems. We also aim to follow additional ideas to improve the detector, such as:

i) Following (Luhn, 1958), we propose *a*) to compute the measure of significance using statistical information derived from lemma frequency and distribution (inside the EDU or the text). *b*) To remove the candidates in all the parentheticals. *c*) To consider only the first EDU which has a verb with a pronoun. *d*) To add information about discourse markers in order to detect multiple central units as in Example (2), which has not any indicator in the following coordinated phrase.

ii) To check the system and the results. For this objective, a qualitative analysis has to be done.

iii) To apply machine learning techniques using the studied indicators as features.

iv) To test our system in the Multilingual RST Treebank (Iruskieta, da Cunha, and Taboada, 2015).[16]

v) To identify the most prominent units from different sections of scientific articles.

vi) To put together the Basque discourse segmenter EusEduSeg[17] (Iruskieta and Zapirain, 2015) and this system to use RST annotation in different tasks such as question answering (Aldabe et al., 2013) and sentiment analysis (San Vicente, Agerri, and Rigau, 2014).

References

Aldabe, Itziar. 2011. Automatic exercise generation based on corpora and natural language processing techniques. Unpublished doctoral dissertation, UPV/EHU, Donostia, Basque Country.

Aldabe, Itziar, Itziar Gonzalez-Dios, Iñigo Lopez-Gazpio, Ion Madrazo, and Montse Maritxalar. 2013. Two approaches to generate questions in basque. *Procesamiento del lenguaje natural*, 51:101–108.

[16]The Multilingual RST Treebank can be consulted at http://ixa2.si.ehu.es/rst/.

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

Alkorta, Jon, Koldo Gojenola, Mikel Iruskieta, and Alicia Perez. 2015. Using discourse topic in basque sentiment analysis. In *SEPLN*.

Burstein, Jill C., Daniel Marcu, Slava Andreyev, and Martin S. Chodorow. 2001. Towards automatic classification of discourse elements in essays. In *Proceedings of the 39th annual Meeting on Association for Computational Linguistics*, pages 98–105. Association for Computational Linguistics.

Egg, Markus and Gisela Redeker. 2010. How complex is discourse structure? In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, page 1619–1623, Valletta, Malta, 19-21 May.

Iruskieta, Mikel. 2014. Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean (a description of pragmatics rhetorical structure and its evaluation in computational linguistic). Phdthesis, Euskal Herriko Unibertsitatea, Donostia. http://ixa2.si.ehu.es/~jibquirm/tesia/tesi_txostena.pdf.

Iruskieta, Mikel, María Jesus Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez, Mikel Lersundi, and Oier 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. http://ixa2.si.ehu.es/diskurtsoa/en/.

Iruskieta, Mikel, Iria da Cunha, and Maite Taboada. 2015. A qualitative comparison method for rhetorical structures: Identifying different discourse structures in multilingual corpora. *Language Resources and Evaluation*, 49:263–309.

Iruskieta, Mikel, Arantza Diaz de Ilarraza, and Mikel Lersundi. 2014. The annotation of the central unit in rhetorical structure trees: A key step in annotating rhetorical relations. In *COLING*, pages 466–475. Dublin City University and ACL, Dublin, Ireland.

Iruskieta, Mikel, Arantza Diaz de Ilarraza, and Mikel 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)*, Cuiaba, Brasil, 24-26 October.

Iruskieta, Mikel and Beñat Zapirain. 2015. Euseduseg: A dependency-based edu segmentation for basque. In *SEPLN*.

Luhn, Hans Peter. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Neto, Joel Larocca, Alexandre D Santos, Celso AA Kaestner, and Alex A Freitas. 2000. Generating text summaries through the relative importance of topics. In *Advances in Artificial Intelligence*. Springer, pages 300–309.

Paice, Chris D. 1980. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In *3rd annual ACM conference on Research and development in information retrieval*, pages 172–191, Cambridge, June. Butterworth and Co.

Pardo, Thiago A. S., Lucia H. M. Rino, and Maria G. V. Nunes. 2003. GistSumm: A summarization tool based on a new extractive method. *Computational Processing of the Portuguese Language*, pages 196–196.

San Vicente, Inaki, Rodrigo Agerri, and German Rigau. 2014. Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. In *14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 88–97. EACL.

Stede, Manfred, 2008. *RST revisited: Disentangling nuclearity*, pages 33–57. 'Subordination' versus 'coordination' in sentence and text. John Benjamins, Amsterdam and Philadelphia.

van Dijk, Teun A. 1980. *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. L. Erlbaum Associates Hillsdale, NJ.