Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 65-72 recibido 18-11-2015 revisado 23-12-2015 aceptado 05-02-2016

Detecting the central units in two different genres and languages: a preliminary study of Brazilian Portuguese and Basque texts

Detección de la unidad central en dos géneros y lenguajes diferentes: un estudio preliminar en portugués brasileño y euskera

Mikel Iruskieta
University of the Basque Country (UPV/EHU)
IXA Group
Sarriena auzoa z/g.
Leioa.
mikel.iruskieta@ehu.eus

Gorka Labaka
University of the Basque Country (UPV/EHU)
IXA Group
Manuel Lardizabal 1.
Donostia.
gorka.labaka@ehu.eus

Juliano Desiderato Antonio
Universidade Estadual de Maringá
Programa de Pós-Graduação em Letras
Maringá - PR - Brasil
jdantonio@uem.br

Abstract: The aim of this paper is to present the development of a rule-based automatic detector which determines the main idea or the most pertinent discourse unit in two different languages such as Basque and Brazilian Portuguese and in two distinct genres such as scientific abstracts and argumentative answers. The central unit (CU) may be of interest to understand texts regarding relational discourse structure and it can be applied to Natural Language Processing (NLP) tasks such as automatic summarization, question-answer systems or sentiment analysis. In the case of argumentative answer genre, the identification of CU is an essential step for an eventual implementation of an automatic evaluator for this genre. The theoretical background which underlies the paper is Mann and Thompson's (1988) Rhetorical Structure Theory (RST), following discourse segmentation and CU annotation. Results show that the CUs in different languages and in different genres are detected automatically with similar results, although there is space for improvement.
Keywords: Central unit, RST, indicators, rules.

Resumen: El objetivo de este trabajo es presentar las mejoras de un detector automático basado en reglas que determina la idea principal o unidad discursiva más pertinente de dos lenguas tan diferentes como el euskera y el portugués de Brasil y en dos géneros muy distintos como son los resúmenes de los artículos científicos y las respuestas argumentativas. La unidad central (CU, por sus siglas en inglés) puede ser de interés para entender los textos partiendo de la estructura discursiva relacional y poderlo aplicar en tareas de Procesamiento del Lenguaje Natural (PLN) tales como resumen automático, sistemas de pregunta-respuesta o análisis de sentimiento. En los textos de respuesta argumentativa, identificar la CU es un paso esencial para un evaluador automático de considere la estructura discursiva de dichos textos. El marco teórico en el que hemos desarrollado el trabajo es la *Rhetorical Structure Theory* (RST) de Mann y Thompson (1988), que parte de la segmentación discursiva y finaliza con la anotación de la unidad central. Los resultados demuestran que las unidades centrales en diferentes lenguas y géneros son detectadas con similares resultados automáticamente, aunque todavía hay espacio para mejora.
Palabras clave: Unidad central, RST, indicadores, reglas.

ISSN 1135-5948

1 Introduction

The development of applications which automatically perform complex linguistic tasks such as summarizing, segmenting, translating and even evaluating texts depends on the linguistic description not only of formal grammar rules, but also on the analysis of discourse structure.

A notion which plays an important role in discourse analysis is the notion of *topic of discourse*. According to van Dijk (1980), language users are able to summarize discourses, expressing the main topics of the summarized discourse. The dutch linguist argues that discourse topics are properties of the global meaning of the text and a necessary feature for the text to be globally coherent. In van Dijk's words, discourses are "organized *around* a semantic 'core' that we intuitively call a theme or topic" (van Dijk, 1980: 41).

In NLP the notion of discourse topic is also very important and the summary of the global meaning of texts has received different tags (Iruskieta et al., 2015): thesis statement (Burstein et al., 2001), central proposition (Pardo, Rino and Nunes, 2003), central subconstituent (Egg and Redeker, 2010), central unit (Stede, 2008). As this paper is developed under the framework of Rhetorical Structure Theory - RST (see Section 2 ahead), we choose Stede's term "central unit" (the most salient node of the rhetorical structure tree).

The detection of the central unit (henceforth CU) is an important key step in the annotation of the relational structure of a text (Iruskieta, Ilarraza and Lersundi, 2014) and can be useful in NLP tasks such as automatic summarization, automatic evaluation, question-answer systems and sentiment analysis. Thus, the aim of this paper is to present the development of a rule-based automatic detector which identifies the CU in two different genres produced in two different languages: scientific abstracts in Basque (henceforth EUS) and argumentative answers in Brazilian Portuguese (henceforth BP).

In RST diagrams, represented as trees (henceforth RS-trees), at least one elementary discourse unit[1] (henceforth EDU) functions as the main nucleus of the tree. It is important to notice that the CU does not function as satellite of any other unit or text span. Two examples of CUs of the corpus are presented below:

(1) ***Lan honetan*** *patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko* ***garrantzitsuen****ak* ***analizatzen*** *ditu****gu****.* [GMB0301]
In this paper we analyze the most important epidemiological, etiopathological, pathological and clinical features of this common oral pathology.

(2) *O* ***segredo*** *do* ***vestibular*** *é sem* ***dúvida*** *o* ***esforço****.* [M21315]
The ***secret*** of ***Vestibular*** is ***without any doubt*** the ***effort***.

Example (1) is from the EUS corpus of scientific abstracts. It was identified by the two annotators of the corpus as the CU of the text. The identification relies on the following indicators: i) '*Lan honetan*' "in this work" in Basque, the demonstrative '*hau*' "this" 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, Ilarraza and Lersundi, 2014); 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.

Example (2) is from the BP corpus of argumentative answers. The analysis of that EDU unveils the indicators used by the annotators of the corpus to identify it as the CU of the text: i) the CU starts with the resumption of the question that was answered by the writers '*Qual o segredo do vestibular: inteligência, esforço ou sorte?*' "What's the secret of Vestibular: intelligence, effort or luck?". Thus, the answer starts as '*O segredo do Vestibular é*' "The secret of Vestibular is"; ii) the noun '*esforço*' "effort" is in compliance with one of the factors suggested in the question; iii) Asseverative epistemic adverbial phrase '*sem dúvida*' "without any doubt" is used by the writers to make their propositions more credible.

It is important to notice that the characteristics of the genre are crucial for the identification of the CU, but the detection has to

[1] EDUs are "minimal building blocks of a discourse tree" (Carlson and Marcu, 2001: 2). In general, clauses are EDUs except for complement and restrictive clauses.

be made based on the elements that constitute the CU.

In order to achieve the goals presented previously, this paper is organized in three more sections. In section 2, we lay out the main tenets of the theory that underlies the paper, the research corpus and the methodology used in the research. Section 3 focuses on the presentation of the system and section 4 sets out the results of the detector. In the final section, conclusions of this study are exhibited.

2 Theoretical framework

RST is a theory which aims at investigating text coherence, especially regarding relations held between parts of text, both in macro and microstructure (Mann and Thompson, 1988). According to Matthiessen (2005), RST emerged from the researches made by a group led by William C. Mann in the beginning of the 1980's, at University of California *Information Sciences Institute*. The group aimed at investigating text organization with the purpose of automatic text generation. Two reputed linguists, Christian Matthiessen and Sandra Thompson, joined the group, which also had the consultancy of Michael Halliday, author-founder of Systemic Functional Grammar. Matthiessen (2005) claims that the group did not imagine that the theory they were creating would arouse so much interest both in Computational Linguistics and in Theoretical Linguistics.

In Linguistics, RST is a framework for the analysis of texts. It is very useful for the description of the superstructure of diverse text genres. Besides that, RST is a prominent theory in Functional Linguistics regarding the investigation of clause combining, describing the relations which are held between clauses in microstructure (Matthiessen and Thompson, 1988).

A relevant aspect of RST is the fact that the theory can be applied to any language and that it can be used to describe almost all text genres, according to Marcu (2000). Many languages have already been annotated using RST: Carlson et al., (2002) annotated manually newspaper articles in English. Taboada and Renkema (2011) annotated, besides newspaper articles, advertisements, letters, magazine articles, scientific papers, book reviews and opinion articles. Stede (2004) annotated newspaper articles in German. Pardo and Seno (2005) annotated texts about computing in Brazilian Portuguese, Cardoso et al., (2011) composed of news texts and Antonio and Cassim (2012) annotated a corpus of spoken discourse. Da Cunha et al., (2011) annotated scientific papers of diverse areas in Spanish. Iruskieta et al., (2013) annotated abstracts of scientific texts in Basque.

Tools for performing automatic tasks have been designed using RST: automatic segmenters for English (Marcu, 2000; Tofiloski and Brooke et al., 2009), for Brazilian Portuguese (Pardo, 2008),[2] for Spanish (Da Cunha and San Juan et al., 2012) and for Basque (Iruskieta and Zapirain, 2015).[3]

Within RST framework, many parsers for automatic discourse analysis have been designed: for example, there are analyzers for Japanese (Sumita and Ono et al., 1992), for English (Corston-Oliver, 1998; Marcu, 2000; Hanneforth and Heintze et al., 2003; Joty and Carenini et al., 2015) and for Brazilian Portuguese (Pardo and Nunes et al., 2004).

A good summary of what has been done about and with RST is available at Taboada and Mann (2006) and there is plenty of information about the theory at http://www.sfu.ca/rst.

1.1 Methodology

The Basque corpus (EUS) used in this paper (see Table 1) consists of abstracts from five specialized domains (medicine, terminology, science, health and life). i) Medical texts include the abstracts of all medical articles written in Basque in the Medical Journal of Bilbao between 2000 and 2008. ii) Texts related to terminology were extracted from the proceedings of the International Conference on Terminology organized in 1997 by UZEI. iii) Scientific articles are papers from the University of the Basque Country's Faculty of Science and Technology Research Conference, which took place in 2008. iv) Health texts include abstracts of papers from 2nd Encounter of Researches of the Health Science organized in 2014 by the Summer Basque University (UEU). v) Life science texts include abstracts of articles from the 1st Encounter of Researches organized in 2010 by the Summer Basque

[2] Senter can be downloaded from http://www.icmc.usp.br/~taspardo/SENTER_Por.zip

[3] The EusEduSeg segmenter for Basque is available at http://ixa2.si.ehu.es/EusEduSeg/EusEduSeg.pl

University (UEU). The Basque corpus (EUS) contains 100 texts, each with its CUs.[4]

The Brazilian Portuguese corpus (BP) (see Table 1) consists also of 100 texts written by candidates for summer 2013 entrance exams[5] at Universidade Estadual de Maringá (UEM). There are excerpts the candidates can base upon to write the texts demanded by the instructions. On Summer 2013 the instructions for argumentative answer were: *As a candidate, write, using up to 15 lines, an argumentative answer to the question "**What is the secret of Vestibular: intelligence, effort or luck?**"*.

A more detailed description is presented in Table 1.

Corpus	Genre	Words	EDUs	CUs
EUS	Abstracts	25,593	2,302	122
BP	Arg. answers	14,285	1,422	116

Table 1: Corpora description: genre and size

According to Swales (1990), scientific abstracts texts follow the IMRaD (introduction, method, results and discussion) structure. The central unit is usually in the introduction part, but sometimes an introductory part is necessary for a better understanding of the main topic. This is represented in RST with the BACKGROUND rhetorical relation.

According to Menegassi (2011), argumentative answer genre belongs to scholar/academic sphere. It is initiated by the resumption of the question followed by the answer to the question, which is the thesis defended by the author. The remainder of the text presents arguments that support the thesis in order to try to convince or persuade the reader.

The size of the corpus for each language studied is similar in size which was used in bibliography (Paice, 1981; Burstein, 2001) for similar aims. For Basque corpus, we have used the Science, medicine and linguistics subcorpora as training (60 texts) and the life and health subcorpora as test data-sets (40 texts). And for BP the first 60 texts were used for training and the last 40 for test.

Both corpora were annotated by two linguists who were familiar with RST and the annotation phases represented in Figure 1 were as follows:

1. Annotators segmented the texts into EDUs manually with RSTTool (O'Donnel, 2000).
2. Annotators determined the CU of each text.
3. The results were evaluated and a segmented gold standard corpus with the annotated CUs was created. The inter-annotator agreement in Basque was 0.796 kappa (for a total of 2440 EDUs). For BP the four annotators identified the same central unit in 75% of the texts (full agreement).
4. The gold standard corpus was annotated automatically with morphosyntactic information and exported to a MySQL database.[6]
5. CU's indicators were manually extracted in each corpus.
6. Heuristics that exploit these CU's indicators were defined for EUS and BP in the training data-set.
7. The results were evaluated against the test data-set of EUS and BP.

3 The system

Our CU identification system is based on the indicators defined in Iruskieta et al., (2015) for Basque and in Antonio (2015) for BP. To do that, each EDU was automatically analyzed and a number of representative features were extracted for each language. Those features include the number of occurrences of each indicator type (from a relevant noun list, verb list, pronouns, demonstratives and bonus word list are used in each EDU), the position of the given EDU into the whole document and the number of words of the title present in the given EDU.[7] Based on those features and using the training corpora for validation, we have defined and tested a number of handcraft heuristics.

[4] Each CU may have more than one EDU, as it can be noticed in Table 1.

[5] The exams are available at http://www.vestibular.uem.br/2013-V/uemV2013p2g1.pdf.

[6] The Basque texts can be found at http://ixa2.si.ehu.es/diskurtsoa/segmentuak.php and the Brazilian Portuguese at http://ixa2.si.ehu.es/rst/pt/segmentuak_multiling.php

[7] Each EUS document contains its own title, but all BP documents share the same title (the questions that the students have to answer '*Qual o segredo do vestibular: inteligência, esforço ou sorte?*').

Those heuristics define the minimum requirements needed to mark an EDU as CU.

Due to the differences in genre and domain between the EUS and BP texts, we have calculated how difficult can be the task of determining the central unit as follows:

- Difficulty = Total of CUs in the data-set / total of EDUs in the data-set.

where the nearer is from 1, the easier it is to determine the CU.

Therefore, in the EUS training data-set the difficulty is 0.063 (78 CUs out of 1236 EDUs), while in BP it is 0.079 (67 CUs out of 846 EDUs). In the EUS test data-set it is 0.041 (44 CUs out of 1066 EDUs), whereas in BP it is 0.085 (49 CUs of 576 EDUs). Looking at these measures, we conclude that detecting a CU in the EUS corpus is more difficult than detecting the CU in BP corpus.

The differences in genre and domain also vary for each language in order to get the best heuristics based on the CU's indicators. For Basque, an EDU has to contain at least two nouns, one noun followed by a determiner or preceded by a pronoun or a verb and has to appear within the first 18 EDUs of the document to be considered a CU. That is, all the features except the bonus words and the words from the title are used. Otherwise, for Brazilian Portuguese, best results are achieved combining only the number of occurrences of words in the title and the nouns and the position of the EDU within the document. Thus, to be considered CU, the EDUs must contain at least three nouns of the list or three words of the title and they have to appear after the question within the second EDU position of the documents (see results of the 'best heuristic' in Table 2).

Alternatively, a numerical method has been used to try to detect CUs. Based on the same numeric features used in the heuristics, we linearly combined them to get an aggregate score used to determine if an EDU is considered a CU (when the score of the EDU is bigger than 1) or not (when the score is smaller than 1). For example, if we defined a weight of 0.3 any noun indicator, any verb indicator and any word in the title and 0.1 for the EDU position (if there is between the first and the second position, and 0 otherwise). An EDU would be marked as a CU if it contained one of each indicator and if there is within the second position (0.3*1+0.3+1+0.3+0.1*1=1) or if it has 4 occurrences of any of the mentioned indicators (0.3*4=1.2). Those weights are manually defined to maximize the results obtained in the training data, and later evaluate unseen examples of the test data (see results of the 'linear comb.' in Table 2).

4 Results

The performance of the heuristics is reported following the standard measures precision, recall and f-score (F1). We calculate each of the measures as follows:

- precision = $correct_{CU}$ / $correct_{CU}$ + $excess_{CU}$
- recall = $correct_{CU}$ / $correct_{CU}$ + $missed_{CU}$
- F1 = 2*precision*recall / precision + recall

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

Table 2 shows the results obtained for Basque:

- The best heuristic considers CU only the EDUs that there are in the position from 2 to 18 and those EDUs that satisfy any of the following constraints: i) two nouns; ii) a noun with a demonstrative pronoun which is within the distance of three words; iii) a word noun with a personal pronoun which is within the distant of three words; and iv) a verb with a auxiliary verb with the first personal pronoun.
- Linear combination considers the following weights: Nouns (*0.2), verbs (*0.2), pronouns (*0.3), auxiliary verbs with the first personal pronoun (*0.2), a combination of a noun with a determiner (*0.8), a combination of a verb with an auxiliary verb with the first personal pronoun (*0.5), a bonus word (*0.525), a title word (*0.05), the EDU position between 2 and 18 (*0.001) and a main verb (*0.1).

And for Brazilian Portuguese:

- The best heuristic considers CU only EDUs that there are in first or second positions and have at least three nouns or three title words.
- Linear combination considers the following weights: nouns (*0.1), a title word (*0.3) and the second EDU position (*0.2).

		Brazilian Portuguese (BP)			Basque (EUS)		
		Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Dev.	**Best heuristic**	0.824	0.627	0.712	0.436	0.519	0.474
	Linear comb.	0.671	0.731	0.700	0.377	0.544	0.446
Test	**Best heuristic**	0.778	0.429	0.553	0.705	0.403	0.512
	Linear comb.	0.535	0.469	0.500	0.280	0.636	0.389

Table 2: Results of the system

Those results from Table 2 show that differences between genres and domains are very clear. For Basque, most of the features are used, while for Brazilian Portuguese only position, nouns and title words are taken in consideration with different weights. Let us underline the biggest differences:

- The title words in argumentative answer texts (some of them are nouns) are a good indicators of the CU, because the students have to argue with the resumption of the question followed by the answer to the question, which is the thesis defended by the author (Menegassi 2011).
- The position of the CU in the document is more restricted in argumentative answer texts than in scientific abstracts. For scientific abstracts the best results were obtained within 2 and 18 and for argumentative answer were within 1 and 2. So it is important to write the CU at the beginning of the argumentative answer texts, while in the scientific abstracts it is between the beginning and the middle, because scientific abstracts need some background information to understand the main topic of the abstract.

5 Conclusions and future works

This paper presents the first study of how the CU can be detected for different languages and different genres following similar rule based heuristics and a linear combination for Basque and Brazilian Portuguese texts. Heuristics and the linear combination were implemented using gold standards extracted from the RST Basque Treebank and Brazilian Portuguese Treebank, which are freely available.[8]

We conclude that the way of indicating the CU is sensible to genre, because studied features or indicators are different and have different weights in its detection. The difficulty of the task is also different depending on the genre. The best heuristic for scientific abstracts is more complex because the task is harder (difficulty of 0.041), whereas for argumentative answers it is 0.085. It is our hypothesis that it is for this reason that we obtained the lower result of 0.041 (test data-set was 0.553 for BP and 0.512 for EUS).

The work carried out will be useful for adding discourse hierarchy information to certain language processing tasks for both languages, such as automatic summarizers, question answering and automatic evaluation of the position and the way of indicating the main idea.

The authors will develop machine learning techniques to improve such promising results and will work with other languages re-utilizing annotated corpora, based on the indicators and heuristics extracted from those corpuses, in similar genres.

In terms of future work, it would be interesting to make a contrastive study of the same genre in Basque and Portuguese. That was not possible for this study because there are not Brazilian Portuguese abstract manually annotated or Basque argumentative texts manually annotated with RST.

[8] The Basque files can be download from http://ixa2.si.ehu.eus/diskurtsoa/fitxategiak.php and the Brazilian Portuguese files from http://ixa2.si.ehu.eus/rst/pt/fitxategiak_multiling.php

References

Antonio, J. D. 2015. Detecting central units in argumentative answer genre: signals that influence annotators' agreement. In *5th Workshop "RST and Discourse Studies"*, in *Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural* (SEPLN 2015), Alicante (España).

Antonio, J.D., and F.T.R. Cassim. 2012. Coherence relations in academic spoken discourse. *Linguistica* 52, pp. 323–336.

Burstein, J.C., D. Marcu, S. Andreyev, and M.S. Chodorow. 2001. Towards automatic classification of discourse elements in essays. In *Proceedings of the 39th annual Meeting on Association for Computational Linguistics*, pp. 98–105. Association for Computational Linguistics.

Cardoso, P.C.F., E.G. Maziero, M.L.C. Jorge, E.M.R. Seno, A. Di Felippo, L.H.M. Rino, M.G.V. Nunes, and T.A.S. Pardo, 2011. CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. *Proceedings of the 3rd RST Brazilian Meeting*, pp. 88–105. Cuiabá/MT, Brasil.

Carlson, L., M.E. Okurowski, and D. Marcu. 2002. *RST Discourse Treebank, LDC2002T07 [Corpus]*. Philadelphia: PA: Linguistic Data Consortium.

Corston-Oliver, S. 1998. Identifying the linguistic correlates of rhetorical relations, *Proceedings of the ACL Workshop on Discourse Relations and Discourse Markers* 1998, pp. 8–14.

Da Cunha, I., E. San Juan, J.M. Torres-Moreno, M. LLoberese, and I. Castellóne. 2012. DiSeg 1.0: The first system for Spanish discourse segmentation. *Expert Systems with Applications*, 39(2), pp. 1671–1678.

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

Da Cunha, I., J.M. Torres-Moreno, and G. Sierra. 2011. On the Development of the RST Spanish Treebank, *5th Linguistic Annotation Workshop (LAW V '11)*, 23 June 2011, Association for Computational Linguistics, pp. 1–10.

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

Hanneforth, T. Heintze, S. and Stede, M. 2003. Rhetorical parsing with underspecification and forests, *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers-Volume 2* 2003, Association for Computational Linguistics, pp. 31–33.

Iruskieta, M. Diaz de Ilarraza, A. Labaka, G. Lersundi, M. 2015. The Detection of Central Units in Basque scientific abstracts. In *5th Workshop "RST and Discourse Studies", in Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN 2015)*, Alicante (España).

Iruskieta, M. Díaz de Ilarraza, A. Lersundi, M. 2014. The annotation of the Central Unit in Rhetorical Structure Trees: A Key Step in Annotating Rhetorical Relations. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 466–475, Dublin, Ireland. August 23-29.

Iruskieta, M. Aranzabe, M.J. Diaz de Ilarraza, A. Gonzalez, I. Lersundi, I. Lopez de Lacalle, O. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations, *4th Workshop RST and Discourse Studies*, Sociedad Brasileira de Computação, Fortaleza, CE, Brasil. October 2013.

Iruskieta M. and Zapirain B. 2015. EusEduSeg: a Dependency-Based EDU Segmenation for Basque. In *Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN 2015)*, Spain. September 2015.

Joty, S. Carenini, G. and Ng, R.T. 2015. CODRA: A Novel Discriminative Framework for Rhetorical Analysis. *Computational Linguistics*, 41(3), pp. 385–435.

Mann, W.C., and S.A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse,* 8(3), pp. 243–281.

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

Matthiessen, C. 2005. Remembering Bill Mann. *Computational Linguistics,* v. 31, n. 2, pp. 161–172.

Matthiessen, C., and S. Thompson. 1988. The structure of discourse and 'subordination'. In: Haiman, J. and Thompson, S. (Eds.) *Clause Combining in Grammar and Discourse*. Amsterdam/Philadelphia: J. Benjamins, pp. 275–329.

Menegassi, R.J. 2011. A Escrita na Formação Docente Inicial: Influências da Iniciação à Pesquisa. *Signum: Estudos da Linguagem*, 14(1), pp. 387–419.

O'Donnell, M. 2000. RSTTool 2.4: a markup tool for Rhetorical Structure Theory. *First International Conference on Natural Language Generation*. pp. 253–256.

Pardo, T.A.S., and M.G.V. Nunes. 2008. On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. *Journal of Theoretical and Applied Computing*, Vol. 15, N. 2, pp. 43–64.

Pardo, T.A.S., and E.R.M. Seno. 2005. Rhetalho: um corpus de referência anotado retoricamente [*Rhetalho: a rhetorically annotated reference corpus*], *Anais do V Encontro de Corpora*, 24-25 November 2005.

Pardo, T.A.S., L.H.M. Rino, and M.G.V. Nunes. 2003. GistSumm: A summarization tool based on a new extractive method. *Computational Processing of the Portuguese Language*, pp. 210–218.

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

Scott, D.R., J. Delin, and A.F. Hartley. 1998. Identifying congruent pragmatic relations in procedural texts. *Languages in Contrast*, 1(1), 45–82.

Stede, M. 2004. The Potsdam commentary corpus, *2004 ACL Workshop on Discourse Annotation*, 25-26 July 2004, Association for Computational Linguistics, pp. 96–102.

Swales, J.M. 1990. Genre analysis: English in academic and research settings. Cambridge, UK: Cambridge University Press.

Sumita, K., K. Ono, T. Chino, and T. Ukita. 1992. A discourse structure analyzer for Japanese text, 1992, ICOT, pp. 1133–1140.

Taboada, M., and W.C. Mann. 2006. Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4), pp. 567–588.

Taboada, M., and J. Renkema. 2011. Discourse Relations Reference Corpus [Corpus]. Simon Fraser University and Tilburg University. Available from http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html.

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

Van Dijk, T. 1980. *Macrostructures: an Interdisciplinary Study of Global Structures in Discourse, Interaction and Cognition*. Lawrence Erlbaum, Hillsdale.