An analysis of textual coherence in academic abstracts written in Portuguese

Vinícius Mourão Alves de Souza and Valéria Delisandra Feltrim
Informatics Departament, State University of Maringá

Abstract

The abstract can be considered as one of the most important sections of an academic work. Along with the title, it is used by researchers to disseminate their research in scientific circles. In this context, it was developed an environment to support the writing of Abstract and Introduction sections called SciPo (Scientific Portuguese). This environment provides writing support by means of critics and suggestions presented to the user with respect to the rhetorical structure identified in text submitted for analysis. Although SciPo provides feedback indicating which parts of the text should be improved, it does not analyze features related to semantics, such as coherence, which is essential to the readability and interpretability of the text. This paper presents a study with regard to the analysis of textual coherence in Abstracts section from Bachelor theses in the area of Computer Science written in Portuguese. The coherence analysis was based on the semantic relationship observed among certain parts of the abstract that constitute its rhetorical structure. This analysis led to the development of additional computational resources to SciPo, which is now able to analyze coherence of the abstracts written in Portuguese.

1. Introduction

The abstract can be considered one of the most important sections of an academic work. Along with the title, it is used by researchers to disseminate their research in scientific circles. According to Feltrim (2004), an abstract should be written very carefully in order to be complete, in terms of necessary information, interesting and informative. In addition, it should present the purpose of the work to the reader and also to encourage the reading of the complete work.

Scientific texts (papers, reports, theses, etc.) have a well-defined structure that can be enunciated as: Introduction, Development, and Conclusion (Huckin & Olsen, 1991; Weissberg & Buker, 1990; Swales, 1990; Trimble, 1985). The Development section may be unfolded in the sections Materials, Methods, and Result, or Materials and Methods, Results, and Discussion, depending on the knowledge area and the nature of the text. In the same way, the Abstract section of these texts also has a well-defined structural scheme that covers essential aspects of the scientific communication, called schematic or rhetorical structure.

In abstracts, types of information and their appearance order are very conventional. Therefore, models have been proposed to guide the writer with respect to the required and optional types of information, as well as their order of presentation. Based on the models of Aluísio et al. (1996); Swales (1990) and Weissberg & Buker

(1990), Feltrim et al. (2003) proposed a model for abstracts in Computer Science area composed of six schematic components arranged in the following order: Background, Gap, Purpose*, Methodology*, Result*, and Conclusion. Each one of these components is composed of sentences that nominate and describe the types of information that are expected in the abstract.

Based on the rhetorical structure model proposed by Feltrim et al. (2003), it was developed an environment to support the writing of Abstract and Introduction sections of Computer Science dissertations written in Portuguese. Such an environment is called SciPo (Scientific Portuguese) (Feltrim, 2004). SciPo provides writing support by means of critics and suggestions presented to the user with respect to the rhetorical structure identified in the text submitted for analysis. Although SciPo provides feedback indicating which parts of the text should be improved, it does not analyse features related to text semantics, such as coherence, which is essential to the readability and interpretability of the text.

Coherence and cohesion are elements responsible for making a set of words or phrases make sense. In this paper, we assume that the coherence concerns the possibility of establishing a logical sense among different sentences of a text. Therefore, it is a principle of interpretability connected to the communication situation and the capacity that the receiver has to calculate the meaning of the text. Thus, it is linked to the text but it does not depend only on the receiver (Kock & Travaglia, 2003; Van Dijk, 1981). Moreover, to make possible the establishment of meaning, it is necessary to use textual elements responsible for the connection between words or sentences, which are pertinent to cohesion (Beaugrande & Dressler, 1981).

Aiming at complementing SciPo's features, particularly for the Abstract section, we have developed computational resources for the automatic detection of certain semantic relations in abstracts, in order to return suggestions related to coherence. Thus, for the construction of the proposed computational resources, it was necessary the collection, annotation and analysis of a specific corpus. Therefore, this paper is aimed at presenting such a corpus, as well as its analysis based on aspects of the Abstract section rhetorical structure and its textual coherence.

This paper is organized as follows: Section 2 presents the steps for collection, annotation and analysis of the corpus and Section 3 presents the conclusion and directions for future work.

2. Corpus Collection, Annotation and Analysis

As this work is in the context of SciPo, the collected corpus was restricted to its knowledge area, which is Computer Science, in particular abstracts from Bachelor theses written in Brazilian Portuguese.

2.1. Collection

All texts that compose the corpus were collected in digital format from three different Brazilian universities departments: Informatics Department of the State University of Maringá (DIN-UEM), Computing Department of the State University of Londrina (DC-UEL), and Informatics Department of the Institute of Physics and Mathematics of

* Considered as a mandatory component, whereas the others are considered as optional.

Federal University of Pelotas (Inf-UFPel). The texts from DIN-UEM were collected directly from the authors or their advisors. Remaining texts were found in the Archiving System and Document Indexing from DC-UEL[1] and the Digital Collection from Library of Science and Technology of UFPel[2]. Even with all texts being available in digital format, it was required a manual effort to extract their contents from the Abstract section, as they were in different file formats.

We collected a total of 385 abstracts of many research areas of Computer Science dated from 1999 to 2009. 205 abstracts out of 385 were collected at DIN-UEM, 98 at DC-UEM and 82 at Inf-UFPel. Table 1 presents the number of abstracts, number of words and average number of words among the identified research areas.

Research Areas	# Abstracts	# Words	Average Words
Database	51	8,720	170.98
Computational Intelligence	70	11,149	159.27
Software Engineering	92	15,482	168.28
Hypermedia	32	5,755	179.84
Digital Systems	36	6,454	179.27
Distributed Systems and Programming Languages	19	4,192	220.63
Graphical Programming	41	6,565	160.12
Computer Networks	44	5,787	131.52
Total	**385**	**64,104**	**171.23**

Table 1. Distribution and total words per area of knowledge in the corpus

2.2. Preliminary Annotation

In order to help the comprehension and manipulation of information contained in the corpus, both by human or computational tools, it is necessary the corpus to be electronically annotated with a set of appropriate tags. In this annotation, we chose to use XML (Extensible Markup Language) to set the annotation scheme.

It was initially used a set of four tags and four attributes to annotate the corpus. Figure 1 presents an abstract in which the preliminary annotation was automatically performed by means of Perl scripts developed for this work.

```
<ABSTRACT id="3">
<Title>Caracterização de Tráfego de Rede</Title>
<P id="0">
<S id="3-0" AZ="B">A caracterização de tráfego de rede tem como objetivos principais:
garantia de qualidade de serviço (QoS), planejar e modelar o tráfego, analisar o
desempenho de redes, analisar a perda de pacotes e estudar o comportamento do usuário.</S>
<S id="3-1" AZ="R">Neste trabalho , serão mostradas algumas técnicas de caracterização de
tráfego como: processos de renovação, processos de Markov, séries temporais e auto-
similaridade .</S>
<S id="3-2" AZ="B">Além destes processos, podem ser utilizados os sketches.</S>
<S id="3-3" AZ="B">Estes possuem a característica de fazer a análise dinâmica e em tempo
real de fluxo de dados, o que é muito importante para detectar anomalias.</S>
</P>
</ABSTRACT>
```

Figure 1. Example of an abstract preliminarily annotated.

1 http://www2.dc.uel.br/nourau
2 http://www.ufpel.tche.br/prg/sisbi/bibct/acervodigital.html

The first tag in the abstract of Figure 1 is represented by "<ABSTRACT> … </ABSTRACT>" and it is aimed at defining the beginning/ending of the abstract. This tag also has the attribute "id" assigned according to the file name (in the case of Figure 1, "3.xml"). The tag "<Title> ... </Title>" has the function of establishing the beginning/ending of the text related to the title of the abstract. The tag "<P> ... </P>" has the function of establishing the beginning/ending for each paragraph and the attribute "id" refers to the position of the paragraph with the count starting at 0. Finally, the tag "<S> ... </S>" has the function of establishing the beginning/ending for each sentence and it has the attributes "id" and "AZ". Each sentence "id" has the value of attribute "id" of the tag "<ABSTRACT>" followed by the position of the sentence in the abstract. The attribute "AZ" refers to the classification assigned to the sentence according to the six components that form the rhetorical structure of an abstract. This attribute is named "AZ" as it follows the idea of "argumentative zones" proposed by Teufel and Moens (2002).

This annotation is called preliminary as it concerns only the rhetorical structure of the abstract. New attributes were added to these tags latter on a second annotation step concerning aspects of coherence (reported in Section 2.4).

2.3. Preliminary Analysis

Based on the components that form the rhetorical structure of an academic abstract proposed by Feltrim (2003), it was carried out the analysis and annotation of each corpus sentence by using the XML tags previously presented. In Table 2 we can observe the frequency of each component in the corpus. The Purpose component is present in most of the analyzed abstracts 97.4% (375 abstracts), followed by the components Background (68.05%), Result (55.32%), Gap (40.51%), Methodology (37.66%), and Conclusion (23.11%).

Component	# Abstracts	Frequency (%)
Background	262	68.05
Gap	156	40.51
Purpose	375	97.40
Methodology	145	37.66
Result	213	55.32
Conclusion	89	23.11

Table 2. Frequency of components observed in the corpus

We obtained a total of 2,293 annotated sentences from a set of 385 abstracts. The distribution of these sentences according to their categories is presented in Table 3.

Component	# Sentences	Distribution (%)
Background	808	35.23
Gap	215	09.38
Purpose	426	18.58
Methodology	273	11.90
Result	451	19.67
Conclusion	120	05.24
Total	**2,293**	**100**

Table 3: Distribution of the components that form the rhetorical structure of analyzed abstracts

We can observe in Table 3 that the sentences classified as Background are present in a larger number in the corpus with a total of 808 sentences, which correspond to 35.23% of all sentences. The sentences classified as Result, Purpose, Methodology, Gap, and Conclusion represents, respectively, 19.67%, 18.58%, 11.90%, 9.38%, and 5.24% of all sentences analyzed. We have observed that although 97.4% of abstracts contain Purpose sentences (Table 2), this kind of sentence is not the highest amount of sentences found in the corpus. This occurs because the 426 sentences classified as Purpose are distributed among 375 abstracts, which gives an average of 1.13 Purpose sentence per abstract. On the other hand, the Background sentences, which appear in a higher number (808), are present in 262 abstracts and with an average of 3.08 sentences per abstract. Thus, we have observed that the component Purpose is present in most abstracts of the corpus (97.4%) with approximately one sentence per abstract, whereas the component Background is present in 68.05% of abstracts of the corpus with an average of 3.08 sentences. Table 4 summarizes the results of this analysis.

Component	# Abstracts	# Sentences	Sentences/Abstract
Background	262	808	3.08
Gap	156	215	1.38
Purpose	375	426	1.13
Methodology	145	273	1.88
Result	213	451	2.12
Conclusion	89	120	1.35
Total	**385**	**2,293**	**5.95**

Table 4. Average amount of sentences per summary, distributed by components of the rhetorical structure

Each sentence of the corpus was semi-automatically annotated with one of the six possible categories provided by the rhetorical structure model. First, each sentence was automatically classified by the statistical classifier AZPort (Feltrim, 2004; Feltrim et al., 2006). As AZPort has an accuracy of proximately 70%, we manually revised the resulting annotated corpus, correcting possible mistakes made by AZPort. This was necessary so that the noise from the automatic annotation of the rhetorical structure does not interfere in the coherence annotation, which was performed as a second step of annotation and is presented in the next section.

2.4. Second Step of Annotation and Analysis

After the first step of annotation in which one of the main tasks is to assign a rhetorical category to each sentence, it was carried out the second step of annotation. In this step, four different types of relationships among sentences have been analysed. Following Higgins et al. (2004), we have named these relations as "dimensions": Dimension Title, Dimension Purpose, Dimension Gap-Background, and Dimension Linearity-break, and they are described in following subsections.

Dimension Title

It verifies the semantic similarity of the analyzed sentence with the title of the abstract. Two values are possible for this dimension: "high", if a high relationship between the sentence and the title is identified and "low" if the relationship is low. Although it is possible that there is a middle ground between the values "high" and "low", we take into account only the two values for this annotation task to make it easier and less

confusing to the annotators. As it is a subjective process, we believe that the inclusion of more values of similarity would make the classification less efficient.

1,050 (46.80%) out of 2,293 sentences were identified as having a low relationship to the title and 1,243 (54.20%) with a high degree of relationship. The distribution of high/low sentences according to the six possible rhetorical categories is

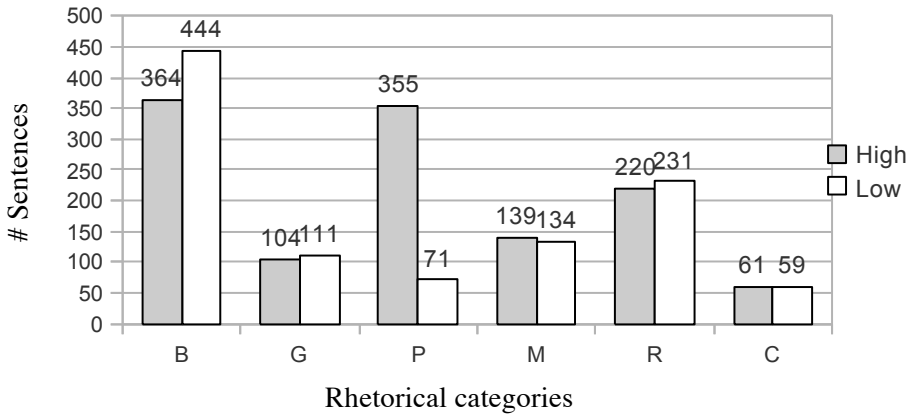

Figure 2. Relationship of rhetorical components with the title

presented in Figure 2.

As it can be observed in Figure 2, there is a strong relationship between sentences classified as Purpose (P) and the title of the abstract, since 355 (83.33%) of these sentences have high relationship to the title, whereas the other categories have an average of 48.79% for high and 51.21% for low sentences. The less related category is Background (B) with 364 (45.05%) of sentences annotated as "high" and 444 (54.95%) of sentences annotated as "low". This behaviour is justifiable as the purpose of Background sentences is to establish the context of the research early in the abstract. In most cases the content of these sentences is only introductory and it does not address topics directly related to the title. Therefore, we believe that the low relationship among these sentences does not characterize a coherence problem. On the other hand, we expect Purpose sentences to indicate the aim of the research briefly describing the main purpose of the work and the same applies to the title. Therefore, it is expected a high level of semantic relationship between these sentences in most cases analysed. With respect to the other rhetorical categories, the semantic relationship with the title is divided, since such relationship depends on the nature of the work and on the way the author explains the problem to be solved, the methodology, the results, and the conclusions. Thus, we assumed that it is not possible to establish a standard level of relationship for these sentences with respect to the title that may indicate a problem of coherence.

A second annotator carried out the annotation of Dimension Title in 20% of the total abstracts analysed by the fist annotator. The second annotator analysed 77 abstracts, summing up a total of 428 sentences, and we used this data to measure the agreement between the two annotators. In terms of the Kappa measure, we obtained K = 0.59, a level of agreement considered as "Good". Details about the agreement/disagreement between annotators can be seen in the confusion matrix of Table 6, where values in the main diagonal represent the number of successes between

the annotators, the values in the extremities represent the totals and other values represent disagreements.

		Annotator 1		
	Class	**High**	**Low**	Total
Annotator 2	**High**	209	59	268
	Low	27	133	160
	Total	236	192	428

Table 6: Confusion matrix for Dimension Title with two Annotators

After the analysis of this dimension, we conclude that it can indicate a possible problem of coherence among important parts of the abstract, such as the sentences that express the main objective of a work and its title. Therefore, we believe that in a coherent abstract, one of its features is that sentences such as those classified in category Purpose have to be semantically related to the title of the abstract.

Dimension Purpose

It verifies the semantic similarity of the target sentence (each sentence of the abstract) with the sentences classified as Purpose. Three values are possible for this dimension: "high", if a high relationship between the sentence and the sentences classified as Purpose is identified, "low" if the relationship is considered low and "N/A" (not applicable) in cases in which the abstract does not have sentences classified as Purpose or when the analysed sentence is Purpose.

Excluding a set of 573 sentences annotated with value "N/A" (divided into 426 Purpose sentences and 147 sentences of abstracts that have no Purpose sentences), it comes a total of 1,720 sentences. 704 (40.93%) out of 1,720 sentences were annotated as having low relationship and 1,016 (59.07%) as having high relationship with the Purpose sentences of their respective abstracts.

The distribution of high/low sentences according to the five rhetorical categories (excluding the category Purpose (P)) is presented in Figure 3.

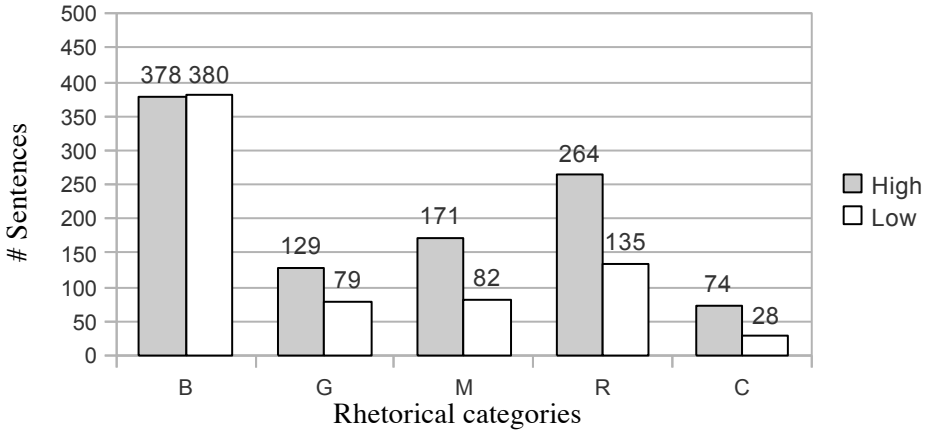

Figure 3. Relationship of rhetorical components with the component Purpose

As shown in Figure 3, it is observed that the categories more related to Purpose are Conclusion (C), Methodology (M) and Result (R) with, respectively, 72.55% (74), 67.59% (171), and 66.17% (264) of high sentences. The less related category is Background (B) with 50.13% (380) of sentences with low relationship with the Purpose of abstract. We have noticed that Background sentences tend to be more related to Gap sentences, which in turn are related to Purpose sentences. We also observed that in most cases sentences of Methodology and Result retakes the content covered in Purpose sentences by using anaphoric expressions. In these cases, the level of semantic relationship is low although these sentences are in fact related to the Purpose.

It is worth notice that during the analysis the annotators experienced difficulties in establishing the relationship between sentences of the categories Methodology and Result with sentences of Purpose. They ascribe these difficulties to the fact that in most cases sentences of Methodology and Result show new terms, such as names of techniques, methods and metrics, making the task difficult even for an experienced annotator.

As in Dimension Title, a second annotator performed the annotation of the Dimension Purpose in 20% of all abstracts analysed by the first annotator. It was annotated 77 abstracts and a total of 317 sentences (once of the total 428 sentences 111 are of the Purpose and not analyzed) to measure the agreement between annotators. In terms of the Kappa measure, we obtained K = 0.594, a level of agreement considered "Good". Details about the agreement/disagreement between the two annotators can be seen in confusion matrix of Table 7, where values in the main diagonal represent the number of successes between the annotators, the values in the extremities represent the totals and the other values represent disagreements.

		Annotator 1		
	Class	**High**	**Low**	Total
Annotator 2	**High**	158	36	194
	Low	26	97	123
	Total	184	133	317

Table 7: Confusion matrix for Dimension Purpose with two Annotators

Based on the results for this dimension, we conclude that it can be used to verify possible problems of coherence among certain parts of the text, such as Conclusion, Methodology, and Result with sentences of the Purpose category, since we have observed a higher relationship between them. Therefore, we believe that in a coherent abstract, sentences such as those classified in categories Methodology, Result, and Conclusion are expected to present a high semantic relationship with Purpose sentences.

Dimension Gap-Background

As mentioned earlier, Background sentences tend to be closely related to Gap sentences then to Purpose ones. Thus, it is expected for the Gap sentences to be related with at least one sentence of Background. Therefore, we understand that the complete absence of relationship between these components can be an indication of a coherence problem.

For each abstract with Gap and Background sentences in the corpus, we have verified the semantic relationship between the sentences of these categories. Each Gap sentence was labelled as "yes" if it is strongly related with some Background sentence; otherwise, it was labelled as "no".

Apart from 32 sentences belonging to abstracts which do not have Gap/Background sentences, 183 sentences were labelled as "yes" or "no" for this dimension. Over this total of sentences, 137 (74.86%) were labelled as "yes" and 46 (24.14%) were labelled as "no".

We have measured the agreement between two human annotators by the Kappa measure over a randomly selected subset of 46 sentences of the corpus and we obtained K = 0.725. Details about the agreement/disagreement between the annotator can be seen in confusion matrix of Table 8 where values in the main diagonal represent the number of successes between the annotators, the values in the extremities represent the totals and the other values represent disagreements.

		Annotator 1		
	Class	**Yes**	**No**	Total
Annotator 2	**Yes**	35	1	36
	No	3	7	10
	Total	38	8	46

Table 8: Confusion matrix for Dimension Gap-Background with two Annotators

Taking into consideration the annotation results for this dimension, we have concluded that the analysis of the Dimension Gap-Background can be used to detect possible coherence problems involving the relationship between the rhetorical components Gap and Background.

Dimension Linearity-break

It verifies the existence of a break of linearity among sentences from the abstract. In this dimension, two values are possible: “yes”, if there is some difficulty in establishing a logical sense among the current sentence with its previous or next sentence, and “no” if the sentence conforms to the text flow. In some cases, it is verified only the relationship with the previous sentence, if the sentence been analysed is the last one of the abstract or with the next sentence, if the sentence been analysed is the first one of the abstract.

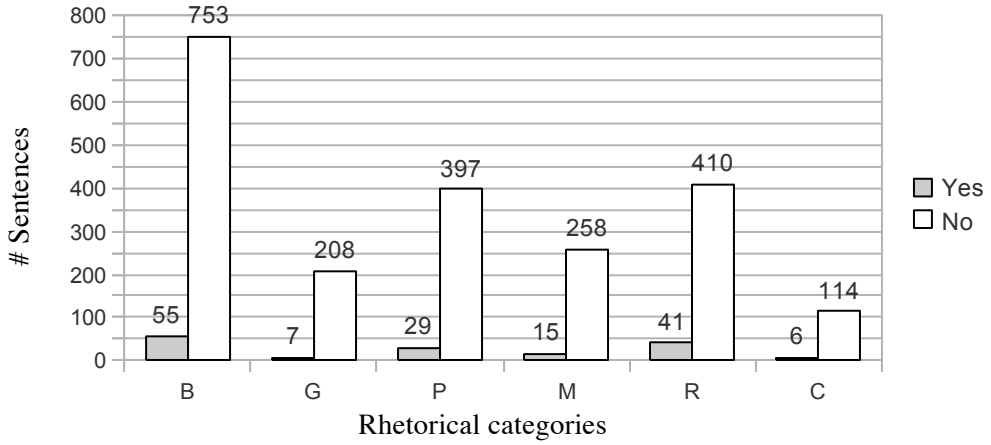

Figure 4. Distribution of rhetorical categories in relation to the Dimension Linearity-break

Only 153 out of 2,293 analysed sentences were identified and annotated with the value “yes” and 2,140 sentences were annotated with the value “no”. It can be seen in Figure 4 that taking the 153 sentences in which the break of linearity were identified, it is most recurrent in terms of proportion in Result (R) sentences, with 41 (9.09%) of sentences with the value “yes”. The sentences of category Gap (G) presents the smallest number of problems concerning linearity, with only seven (3.26%) occurrences of “yes”. We have noticed in this dimension that a significant break in the flow of text among adjacent sentences rarely occur, since the identified cases of “yes” sentences represent only 7.14% of the total analysed sentences. Furthermore, we have observed that in some of these cases the sentence is “connected” to another part of the text that is not an adjacent sentence.

Due to the low number of identified “yes” sentences, the level of agreement between annotators was not calculated. Moreover, it would require a larger number of texts than the used in other dimensions, in order to have a significant number of “yes” sentences identified by both annotators.

Even with a low number of occurrences of “yes” sentences in this dimension, we understand that its analysis is important in the task of identifying coherence problems, especially at a local level of analysis. Since our corpus is composed of abstracts of Bachelor theses, it is natural the occurrence of breaks in linearity to be rare. We believe that this kind of problem may be more frequently found in texts with serious writing problems and this is not the case of the corpus analysed. Thus, we believe the Dimension Linearity-break to be valid if applied to a corpus with a higher level of “noise”, such as texts generated by automatic summarization.

3. Conclusions

The main purpose of this work is to propose four dimensions of analysis concerning textual coherence in academic abstracts. These dimensions take into consideration the rhetorical structure of abstracts, as proposed by Feltrim et al (2003). A corpus of 385 abstracts from Bachelor theses written in Brazilian Portuguese were analysed according to the proposed dimensions and the results for each of them were presented.

By taking into account the manual analysis performed on the corpus, we observed that from the four proposed dimensions, at least three can be automated by means of computational resources: Dimension Title, Dimension Purpose and Dimension Gap-Background. In these three dimensions we have observed the existence of patterns concerning the rhetorical structure and aspects of coherence, differently from the fourth dimension, Dimension Linearity-break.

We observed in Dimension Title that the sentences with a higher semantic relationship with the title of an abstract were the sentences of Purpose category. Moreover, we observed problems of coherence in abstracts in which the relationship of the title and the sentences of the Purpose are low. Both categories of sentences, title and Purpose, summarize the main purpose of the work, each in its proportion and, therefore, a high relationship between these sentences is expected to led to a greater level of coherence.

In Dimension Purpose we have analysed the relationship of different categories of sentences with Purpose sentences. In particular, the sentences of the rhetorical categories Conclusion, Methodology and Result have showed higher levels of relationship with Purpose. In fact, it was possible to verify that the low relationship among these categories sentences makes the abstract more difficult to be interpreted and, thus, less coherent. The high relationship observed between Conclusion sentences and Purpose sentences in most cases can be attributed to the recovery of the content covered in Purpose sentences for finalizing the text, which leads the reader to make associations between these sentences. The low relationship observed between Background sentences and Purpose ones can be explained by the introductory content of Background sentences and its relation to Gap sentences.

Given the observations that the low relationship between Gap and Purpose does not represents an problem of coherence, we have analysed in Dimension Gap-Background the relationship between sentences of the rhetorical category Gap with at least one sentence of Background. In fact, it was possible to verify that the absence of relationship between these categories sentences makes the abstract less coherent.

On the other hand, in Dimension Linearity-break a few cases have been identified and they do not constitute a pattern in this corpus. For this corpus, we have observed that only in rare occasions we found sentences that are not related to the previous or next sentence. Moreover, in some cases the sentence is justified in another part of the text which is not an adjacent sentence. Thus, this dimension cannot be used as an indicator of problems related to coherence in academic abstracts, as previously proposed. However, we believe that this dimension can be used in the coherence analysis of other corpora.

As mentioned earlier, three of the four proposed dimensions have been automated by using machine learning algorithms and features automatically extracted of

the text. The results regarding the performance of each dimension classifier are presented in Souza and Feltrim (2011). As future work, we expect to expand our coherence analysis to other sections of academic texts, such as Introduction or Conclusion sections.

Acknowledgements

We would like to thank CAPES for the financial support as well as the annotators for their invaluable work.

References

Aluísio, S.; Barcelos, I.; Sampaio, J.; Oliveira Jr., O. 2001. How to learn the many unwritten "rules of the game" of the academic discourse: A hybrid approach based on critiques and cases. In: *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, p. 257-260.

Beaugrande, R.; Dressler, W. 1981. *Introduction to text linguistics*. New York, NY: Longman Publisher Group, 288 p.

Feltrim, V.; Aluísio, S.; Nunes, M. 2003. Analysis of the rhetorical structure of computer science abstracts in Portuguese. In: *Proceedings of Corpus Linguistics,* p. 212-218.

Feltrim, V. D. 2004. Uma abordagem baseada em corpus e em sistemas de crítica para a construção de ambientes web de auxílio à escrita acadêmica em português. *Thesis*, ICMC – USP, São Carlos – SP.

Feltrim, V. D.; Teufel, S.; Nunes, M.; Aluísio, S. 2006. Argumentative zoning applied to critiquing novices scientific abstracts. *Computing Attitude and Affect in Text: Theory and Applications,* p. 233–246.

Higgins, D.; Burstein, J.; Marcu, D.; Gentile, C. 2004. Evaluating multiple aspects of coherence in student essays. In: *Human Language Technologies: The 2004 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Huckin, T. N.; Olsen, L. A. 1991. *Technical Writing and Professional Communication For Nonnative Speakers of English*. New York, USA: McGraw-Hill.

Kock, I.; Travaglia, L. 2003. *A coerência textual*. São Paulo, Brasil: Contexto.

Souza, V.A.M; Feltrim, V.D. 2011. Automatic Analysis of Semantic Coherence in Academic Abstracts Written in Portuguese. To appear in: *Proceedings of The 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, Chiang Mai, Thailand.

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

Teufel, S. and Moens, M. 2002. Summarising Scientific Articles – Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, 28 (4), 409-446.

Trimble, L. 1985. *English for science and technology: a discourse approach*. Cambridge, UK: Cambridge University Press.

Van Dijk, T. 1981. *Studies in the pragmatics of discourse*. The Hague/Berlin: Mouton.

Weissberg, R.; Buker, S. 1990. *Writing up research: Experimental research report writing for students of english*. Prentice Hall Regents.