Analysis of the rhetorical structure of computer science abstracts in Portuguese

Valéria D. Feltrim
Sandra M. Aluísio
Maria das Graças V. Nunes
NILC – Computational Linguistics Group
ICMC – University of São Paulo
C.P.668 São Carlos, SP, Brazil
{vfeltrim|sandra|gracan}@icmc.usp.br

1. Introduction

It is widely acknowledged that academic writing is a complex task even for native speakers, since it involves the complexities of the writing process as well as those specific to the academic genre[1] (Sharples and Pemberton 1992). To make it worse, especially to novice writers, often the demands of the academic genre are not clear enough. A number of writing tools have been described in the literature whose ultimate goal is to improve the quality of academic English texts produced by novice and/or non-native writers, e.g. WE (Smith and Lansman 1988), Writer's Assistant (Sharples et al. 1994), Composer (Pemberton et al. 1996), Academic Writer (Broady and Shurville 2000), Abstract Helper (Narita 2000), and Amadeus (Aluísio et al. 2001). These tools provide help during different stages of the writing process, from the generation and organization of ideas to post-processing tasks. For Portuguese, on the other hand, one can only find post-processing tools, such as spell-checkers, grammar checkers and online dictionaries and thesauri.

With the ultimate goal of aiding academic writing, the project SciPo, currently being developed at NILC[2], aims at analyzing the rhetoric structure of Portuguese academic texts — in terms of schematic structure, rhetorical strategies and lexical patterns — to derive models for supporting the creation and evaluation of computational writing tools. To make the project feasible, the analysis has focused on specific sections of theses in Computer Science — abstract, introduction and conclusion, which are the most studied in the literature (Swales 1990, Weissberg 1990, Liddy 1991, Santos 1996) and also have been pointed out as the most difficult ones in a questionnaire applied to graduated students. The choice for this kind of text was made mainly for three reasons: theses having to be written in Portuguese, unlike research articles, which are preferably written in English; the high standardization of Computer Science texts, as in other scientific research areas; and SciPo's developers' familiarity with the Computer Science domain.

A corpus of fifty-two (52) Computer Science theses was compiled and has been used both in the analysis of writing patterns specific to the focused research community and in the identification of their main writing problems. The whole corpus analysis is being carried out in a set of stages, the first of which has been accomplished, namely the analysis of the abstracts. This paper presents the results of this first stage and discusses the features of abstracts at a macro level of textual organization as well as comments on its annotation process.

The next sections present the methodology used for the annotation and analysis of the abstracts (Section 2) as well as the results obtained (Section 3). Finally, features for a writing tool to assist novice academic writers are put forth based on what has been observed in the corpus (Section 4).

2. Corpus annotation

The annotation of the corpus was carried out manually in each of the 52 abstracts, consisting of three levels of analysis: identification of structural components, identification of rhetorical strategies used for the realization of each component, and identification of lexical patterns (similar to the "formulaic expressions" used by Teufel et al. 1999) that can be used as lexical clues of the argumentative role of a sentence. These three levels were focused because they reflect the levels of help that we intend to give in a writing tool, namely: how to organize the text (structural components), how to realize each part of

[1] We call "academic genre" the one employed in published academic works (papers, theses, technical reports, etc) of a research community.

[2] NILC – a Brazilian Computational Linguistics Research Group (www.nilc.icmc.usp.br)

the text (rhetorical strategies) and what are the common expressions used in each part of the text according to the selected strategies (lexical patterns).

In order to render the annotation more reliable, we had four different human annotators annotate our whole corpus of abstracts. Given a scheme of 6 structural components, with 3 rhetorical strategies each, the annotators were asked to identify text fragments corresponding to those components and strategies, not leaving unclassified fragments. We actually had to perform the annotation of our corpus several times, since the first experiments showed low agreement between annotators. This may have been caused mainly by four factors: (1) the initial inadequacy of the annotation scheme, (2) the lack of familiarity of the annotators with the annotation scheme, (3) the nature of our corpus and (4) the subjective nature of this kind of annotation.

Although the first versions of the annotation scheme had been based on models accepted in the literature, like Swales's model, the corpus presented some peculiarities that did not fit in the model. Additionally, the actual meaning of its elements was not thoroughly stated, making room for misinterpretation. Finally, as the annotators were initially not very familiar with the annotation scheme, disagreement between them followed naturally. Another factor of difficulty was the nature of our corpus. Although it is composed of academic theses, the texts probably were not submitted to such rigorous revision as published articles usually are. Many abstracts even presented passages that could not be classified within any category. This made the annotation process still more difficult and also required a critical view from the annotators. In addition, working with Portuguese brought the difficulty of dealing with long sentences. The average number of words per sentence in our corpus is 26.8 while in a corpus of English abstracts[3] it is 23. So, it is common to find different components of the annotation scheme in one sentence, sometimes mixed. The mixed components found in the corpus will be discussed in Section 3, and the major writing problems observed will be described in Section 4.

1 Setting
S1 Arguing about the topic prominence
S2 Familiarizing terms, objects, or processes
S3 Introducing the research topic from the research area

2 Gap
G1 Citing problems/difficulties
G2 Citing needs/requirements
G3 Citing the absence of previous research

3 Purpose
P1 Indicating the main purpose
P2 Specifying the purpose
P3 Introducing more purposes

4 Methodology
M1 Listing criteria or conditions
M2 Citing/Describing materials and methods
M3 Justifying choices for methods and materials

5 Results
R1 Describing the artefact
R2 Presenting/Indicating results
R3 Commenting/discussing on the results

6 Conclusion
C1 Presenting conclusions
C2 Presenting contributions/value of research
C3 Presenting recommendation

Figure 1. Overview of the designed annotation scheme

As we could not change the nature of our corpus, we tried to minimize the subjectivity of the task by focusing on better defining the annotation scheme. Since the annotators took part in this process, they automatically became more familiar with the scheme and likely to master it and agree on its usage. After these adjustments, we managed to reach a stable scheme and higher inter-annotator consistency. As a starting point, the annotation scheme was derived from three models: Swales' CARS (1990) and those by Weissberg (1990) and by Aluísio and Oliveira Jr (1996), the latter modeling introductions of experimental research papers written in English. Although these works deal with

[3] In order to calculate this average, we took 54 abstracts from the Computation and Language (cmp-lg) corpus. The cmp-lg corpus is composed by scientific papers which appeared in Association for Computational Linguistics (ACL) sponsored conferences.

introduction sections, the basic structure of their models could also be applied to abstracts. So, during the first annotation experiment, the scheme was modified to accommodate all the argumentative roles found in the corpus. The components and rhetorical strategies which compose our annotation scheme are similar to the ones presented by those authors, especially to Aluísio and Oliveira Jr's. The final version of our annotation scheme is presented in Figure 1.

Such similarity had been expected and shows that, despite the heterogeneity of the corpus, it exhibits predictable rhetorical patterns of scientific argumentation. The major deviation from the traditional structure pattern was found in sentences reporting results. Many of them focus on the resulting product (mainly computational systems) instead of on the corroboration of the initial hypotheses of the underlying research issue. To accommodate this particularity, we included the new rhetorical strategy *Describing the artifact*. We believe that this "non-standard" strategy to report results stems from the technological nature of Computer Science, in which it is common to emphasize the artifact (i.e. a piece of software or a method) developed during the research.

3. What has been found in the corpus

The texts in our corpus were collected from online theses repositories and date from 1994 to 2000, comprising 49 MSc theses and three PhD theses, most of them written by students from our Computer Science Department. Only 2 texts were written by students from other Brazilian universities. Texts in the corpus span several Computer Science sub-areas. So, we divided them into 7 research topics: database systems, computational intelligence, software engineering, hypermedia, digital systems, distributed systems, and graphical computation and image processing. Figure 2 presents the number of abstracts classified in each research topic as well as the total and average number of words per abstract.

Research Topic	Abstracts	Number of words	
		Total	Average
Database Systems	3	1,006	335.3
Computational Intelligence	7	1,452	207.4
Software Engineering	16	3,202	200.1
Hypermedia	12	2,097	174.7
Digital Systems	1	171	171
Distributed Systems	12	1,962	163.5
Graphical Computational and Image Processing	1	94	94
Total	*52*	*9,984*	*192*

Figure 2. Corpus distribution across Computer Science research topics

Considering the components of the annotation scheme presented in the previous section (Figure 1), 96.2% of the abstracts are classified as strict subsets of that scheme, with some repetitions of components. Only 3.8% of the abstracts contain all the six components, in one of the following sequences: [S G P M C R C][4] or [S G P C M R], neither being in the order recommended by the scheme. Considering the number of different components observed in each abstract (Figure 3), 50% of the abstracts present 5-4 components, 44.3% present 3-2 components and 1.9% present only one component (*Purpose*). As mentioned earlier, repetition of components is very common in the corpus and some patterns[5] have been identified. The most frequent pattern is the repetition of *Setting* followed by *Gap*, i.e. $(SG)^+$, present in 25% of the abstracts. Others repetition patterns have been observed, such as *Methodology* followed by *Result*, i.e. $(MR)^+$, and *Result* followed by *Conclusion*, i.e. $(RC)^+$, but with much lower frequency.

Number of components	Frequency
6	3.8%
5-4	50%
3-2	44.3%
1	1.9%

Figure 3. Numbers of components per abstract

[4] In this section, the letters S, G, P, M, R, C stand for the components of the annotation scheme: **S**etting, **G**ap, **P**urpose, **M**ethodology, **R**esult and **C**onclusion.

[5] We use regular expressions to represent repetition and structure patterns.

Regarding ordering of components, we also observed some patterns, especially involving the components at the beginning of the abstracts. The pattern *Setting* followed by *Gap*, with repetition or not, followed by *Purpose*, i.e. $((SG)^{+}|(GS))P$, appears in 30.7% of the corpus. Instances[6] of this pattern are [S G S G P *R*] and [G S P *R M*]. Another frequent pattern is *Setting* followed by *Purpose*, followed or not by *Methodology* or *Result*, i.e. SP[M|R], which is observed in 21.1% of the corpus. Instances are [S P], [S P R *C*] and [S P M *R*]. The pattern *Purpose* followed by *Result*, with repetition or not, followed or not by *Methodology* or *Conclusion*, i.e. $(PR)^{+}[M|C]$, and the pattern *Purpose* followed by *Methodology*, followed or not by other *Purpose* or *Result* or *Setting* PM[P|R|S] appear, respectively, in 19.2% and 13.4% of the abstracts. Other combinations of components appear in the corpus with a very low frequency, so they were not considered here. A summary of the most frequent patterns is presented in Figure 4.

Pattern	Frequency
$((SG)^{+}\|(GS))P$	30.7%
$SP(M\|R)^{*}$	21.1%
$(PR)^{+}[M\|C]$	19.2%
PM[P|R|S]	13.4%

Figure 4. Observed patterns frequency

The ordering patterns identified show that the order adopted in our annotation scheme is reasonable, since 51.8% of the corpus start with components *Setting* or *Gap* followed by *Purpose*, despite order varying greatly from middle to final components. In fact, we observed a lack of clarity when analysing these components, especially *Methodology* and *Result*. They usually occur mixed with each other or, even more often, with *Purpose*. The frequency of these cases is shown in Figure 5. The most frequent case of mixing occurs between *Result* and *Purpose*, which appears in 38.5% of the corpus. This means that, in 38.5% of the abstracts, the passage annotated as *Indicating the main purpose* also includes traces of *Result*. Also, in 9.6% of the corpus, *Purpose* presents traces of both *Result* and *Methodology*. The high frequency of *Result* mixed with *Purpose* may be viewed as another evidence of the importance given to the resulting product of the research – the artifact – as previously noticed in Section 2. We believe that such importance is the reason why writers focus on the produced artifact inside *Purpose*, since it is a central component with high relevance for the abstract. Still considering the observed patterns, we noticed that many texts present not only similar structure but even the same lexical patterns, mainly when written by students who had the same advisor. Such similarity was also observed in texts written by students from the same research group. This suggests that students tend to use texts produced inside their work group as models when writing their own texts and also shows the great influence of the advisor on the student text.

Mixed Components	Frequency
Result mixed with *Purpose*	38.5%
Methodology mixed with *Purpose*	19.2%
Result and Methodology with *Purpose*	9.6%
Methodology mixed with *Result*	7.7%

Figure 5. Frequency of attached components

The frequency of each component of the annotation scheme in the corpus is presented in Figure 6. The numbers show lower frequencies for *Setting*, *Gap* and *Conclusion* in contrast with components *Purpose*, *Methodology* and *Result*. This is in agreement with the results reported by Santos (1996) and Motta-Roth and Hendes (1998). These authors have also observed the optional character of using initial and final components in abstracts and a tendency to a more frequent use of middle components, especially *Purpose*, which appears in all the analysed abstracts. This strongly suggests that writers see *Purpose* as the most important kind of information that an abstract should provide to the reader, followed by *Result* and *Methodology*. Although, the frequency of *Setting* in our corpus also suggests writers' concern with contextualizing their research, which was not observed by Motta-Roth and Hendes. We believe that this difference is due to the nature of the two corpora. As our corpus is composed by thesis abstracts, the writers were not faced with the constraint of a limit of words, which generally hold for journal abstracts, and so they are free to make extensive contextualization.

[6] All the presented structure examples are taken from the corpus.

Component	Frequency
Setting	55.7%
Gap	42.3%
Purpose	100%
Methodology	63.4%
Result	67.3%
Conclusion	30.7%

Figure 6. Distribution of components

Another aspect related to the nature of our corpus is the high frequency of the rhetorical strategy *Presenting contributions/value of research*. As can be viewed in Figure 6, component *Conclusion* occurs in 30.7% of the corpus. Of these occurrences, 68.8% use the referred rhetorical strategy (or 21.2% of the corpus occurrences). Considering that is very important in MSc/PhD theses to emphasize the contributions to a research area, it seems natural for writers to use that conclusion strategy in the abstracts with some frequency.

4. Observed writing problems

In this paper, when we refer to writing problems we mean deviations from the traditional structure model that can compromise the communicative objective of the abstract. As was previously commented in Section 2, the texts in our corpus probably have not been through such rigorous revision, even though the texts are MSc and PhD theses. So, it is not surprising that the abstracts presented some problems, not only general, like grammatical mistakes, but also specific to the academic genre. Moreover, as the major part of the corpus is MSc theses, we can also attribute the observed problems to the inexperience of the writers. We will not comment on the superficial problems found in the corpus, since these are not our object of study in this paper. Thus, we will only focus on problems that are specific to the academic genre.

The major problems observed in the abstracts were misuse of lexical patterns and verbal tenses, inefficient organization and inappropriate emphasis on some specific components. We regard as misuse of lexical patterns those cases in which the writer uses expressions in a component which are specific to other components. An example is the use of *however*, which is a lexical clue to component *Gap*, in a *Setting* sentence and not with its correct contrasting value. Other examples are expressions of the type "*A and B are also presented.*", which can be seen as an indication of the rhetorical strategy *Introducing more purposes*, used in sentences reporting results in an indicative way. Furthermore, the writers sometimes do not use the appropriate verbal tense, especially in sentences reporting results. It is recommended by the literature (e.g. Weissberg 1990) to use the past tense when reporting findings. This gives the reader the notion that the results were obtained by the reported research and not in a previous one. However, it is common to find sentences in our corpus reporting findings in the present tense. Due to this, sometimes we could not determine quite whether the writer was reporting his own results or commenting on someone else's results (as in literature review). Misuse of lexical patterns and verbal tenses surely demanded great effort from the annotators when it came to interpreting the texts to identify their components.

Regarding inefficient organization, we observed some abstracts in which the writer mixed components in such a way as to confuse readers. An example of this problem can be observed in the structure [P M S G P]. The writer initiated by indicating the main purpose (first P) and then described the methodology used to accomplish that purpose (M). After that, a more natural move would be to present results; however, the writer used a *Setting* component, followed by a *Gap* (SG), in order to lead the reader into the further detailing of the previously stated purpose and the introduction of yet other purposes. The presence of *Setting* and *Gap* in the middle of the abstract, separating the main purpose from its detailing, confuses the reader, who may lose track of the main purpose of the related research. Also, the sequence *Methodology-Setting* disrupts the cohesion of the text, causing the reader to feel that "something is missing".

A different case of inefficient organization is inappropriate emphasis on some components. In our corpus such emphasis was observed mainly in contextualization (Setting + Gap). One such example is an abstract that has the structure [S G S G P] and 160 words. A *Purpose* component, which also has traces of *Result* and *Methodology*, corresponds to 16.9% of the abstract. The rest of it (83.1%) is dedicated to the contextualization of the problem, $(SG)^+$. Taking into account that these writers had no limit of words to write their abstracts, we can consider such an abstract little informative. It would be desirable to find more information about the reported research in addition to such a relatively extensive contextualization.

Though less frequent, three further problems were found in the corpus, namely: (1) passages that could not be classified as a component of the annotation scheme. This also took the annotators some time, as previously commented in Section 2; (2) passages indicating obvious information, like "*this work also presents a literature review about the focused subject"*. It is quite sure that an MSc/PhD thesis will have a literature review on the main focused subjects, fairly obviating that piece of information; (3) exactly one abstract presenting the outline of the theses. This kind of information is common in introductions, but not in abstracts.

4. Conclusions

In this paper we have reported on the annotation and analysis of a corpus of thesis abstracts in Computer Science, based on an annotation scheme designed specially for this project. We have argued that this scheme is stable and reasonable on the basis of the results of the corpus analysis. Furthermore, we have discussed structuring patterns and some particularities of the corpus, such as repetition of components and levels of relevance given to each component, and identified writing problems, which are deviations from what is prescribed as characteristic of a good academic abstract.

As we mentioned in the introduction, the presented corpus analysis is part of a project which aims at deriving models for computational writing tools specific to the academic genre in Portuguese. Based on what was observed in the corpus, including its problems, we are developing computational tools for aiding especially novice writers. These tools are being implemented as Web-based applications featuring (1) prescriptive guidelines related to the academic genre and (2) repositories of good and bad examples of structure, writing strategies and lexical patterns. The example repository will have abstracts, introductions and conclusions. Another tool is meant to be a critiquing system capable of giving advice/criticism on structure organization, verbal tenses and level of emphasis given to each component. As a result, we expect to help writers overcome their difficulties related to the academic genre. As future work, we intend to extend our analysis to introductions and conclusions, implement stable prototypes of these tools, and evaluate them with real users, e.g. graduate Computer Science students.

6. References

Aluísio S M, 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*, Madison/Wisconsin, pp 257-260.

Aluisio S M, Oliveira Jr. O N 1996 A Detailed Schematic Structure of Research Papers Introductions: An Application in Support-Writing Tools. *Revista de la Sociedad Espanyola para el Procesamiento del Lenguaje Natural*, 19: 141-147. Also available in <http://www.cica.es/sepln96/sepln96.html>

Broady E, Shurville S 2000 Developing Academic Writer: Designing a Writing Environment for Novice Academic Writers. In E. Broady (ed.), *Second Language Writing in a Computer Environment*, CILT, London, pp 131-151.

Liddy E D 1991 The Discourse-Level Structure of Empirical Abstracts: An Exploratory Study. *Information Processing & Management*, 27(1): 55-81.

Motta-Roth D, Hendges G 1998 Uma Análise Transdiciplinar do Gênero Abstract. *Revista Intercâmbio*, VII: 125-134.

Narita M 2000 Corpus-based English Language Assistant to Japanese Software Engineers. In *Proceedings of MT-2000 Machine Translation and Multilingual Applications in the New Millennium*. pp 24-1 – 24-8.

Pemberton L, Shurville S, Hartley A 1996 Motivating the Design of a Computer Assisted Environment for Writers in a Second Language. In *Proceedings of CALICE'96*, pp 141-148.

Santos M B 1996 The textual organization of research paper abstracts. *Text* 16(4): 481-99.

Sharples M, Pemberton L 1992 Representing writing: external representations and the writing process. In P.O. Holt and N. Williams (eds.) *Computers and Writing: State of the Art.* Intellect, Oxford, pp 319-336.

Sharples M, Goodlet J, Clutterbuck A 1994 A comparison of algorithms for hypertext notes network linearization. *International Journal of Human-Computer Studies* 40(4): 727-752.

Smith J B, Lansman M 1988 *A Cognitive Basis for a Computer Writing Environment*. Technical Report, n.87-032, Chapel Hill.

Swales J M 1990 *Genre Analysis: English in Academic and Research Settings*. Cambridge applied linguistics series.

Teufel S, Carletta J, Moens M 1999 An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of EACL 1999*.

Weissberg R, Buker S 1990 *Writing up Research: Experimental Research Report Writing for Students of English*. Prentice Hall.