# Rhetorical Browzing in Journalistic Texts: Preliminary Investigations

Patrice Enjalbert, Stéphane Ferrari, Alexandre Labadié

# Rhetorical Browzing in Journalistic Texts: Preliminary Investigations

Patrice Enjalbert, Alexandre Labadié, Stéphane Ferrari
Laboratoire GREYC
Université de Caen & CNRS
Bd Maréchal Juin - BP 5186 F
14032 Caen Cedex, France
FirstName.Name@unicaen.fr

***Abstract*—The work presented in this paper concerns discourse structure analysis and its applications to intra- and inter-document search. In a typical application, which could be called "rhetorical browsing", the system will provide assistance to a journal reader in order to focus on texts and passages presenting certain *kind* of information and comments, according to his/her current interest: may be raw information, possibly with chronological dimension, or on contrary analyses, recommendations, debates, etc.. The discourse model can be related to Swales's "discourse moves" and the derived "argumentative zoning" procedures for scientific documents. However due to the nature of the considered texts, zones are defined in more "generalist" terms, following the classic Narration-Description-Argumentation-Prescription typology and especially C. Smith's notion of "discourse modes". The paper presents some preliminary steps performed in order to test the feasibility of the project. First of all, in order to ground our research on firm observations, we decided to build a corpus of journalistic texts, annotated according to the discourse model in view. Quantified results concerning the organization of discourse modes within texts could be obtained thanks to these annotations. In a second step, an experimental procedure for automatic tagging of text passages according to discourse modes has been designed, implemented and tested on the corpus.**

## I. Introduction

ONE can currently observe an increasing interest for discourse structure analysis in the NLP community, both for applicative purposes (improvement of document indexation, summarization, document browsing, passage extraction...) and corpus-based linguistic studies. A very popular approach tries to capture text organization in terms of successive "homogeneous" blocks, representing the succession of "topics" addressed in the text. This so-called *thematic segmentation* has received many implementations and experimentations, in the line of Hearst's *Text Tiling* [1].

*Rhetorical zoning* is a less represented but developing matter. Notably, a number of on-going works are based on Swales' notion of "discourse moves" [2]. Attempts to automatically discover such structures by means of machine learning techniques notably count the pioneer work of [3] for scientific texts, and extensions to other kinds of texts as in [4]. In order to adapt these ideas to our journalistic corpus, we consider a refinement of the Descriptive-Argumentative-Narrative-Prescription model considered (with many variants) in literary studies [5], [6], [7]. According to this model, texts or passages of texts can be labeled by such a *discourse* (or *rhetoric*) *mode*.

Our interest is strongly related with practical concerns. As news readers we observe that, from one reading to another, we may be interested in a different kind of content: maybe raw information, with possibly strong chronological aspects, or on contrary analyses and explanations, recommendations, etc. And not only different papers will match our expectations, but even, especially in long articles, specific passages in them. Hence an interesting consequence of our work would be inter- and intra- document browsing, according to rhetorical and not only topical criteria.

In order to ground our research on firm observations, we decided to build a corpus annotated according to the discourse model in view. The corpus is composed of in-depth articles in economy and politics from the French newspaper *Le Monde*. The annotation task consisted in a labeling of texts passages with a selected set of discourse modes. Quantified results concerning the organization of discourse modes within texts could be obtained thanks to these annotations. In a second step, an experimental procedure for automatic tagging of text passages according to discourse modes has been designed, implemented and tested on the corpus.

The paper is organized as follows. We first describe the corpus, the discourse model, and the annotation procedure. Quantified results concerning the organization of discourse modes within texts are then presented, completed by the description of the automated tagging procedure.

## II. Corpus, model and annotation procedure

### A. Texts, annotators and tools

The corpus in view is composed of journalistic texts from *Le Monde*, year 1994. This choice is due both to applicative goals and to the linguistic quality of the journal. We randomly selected 30 texts (mainly in politics and economy) of different sizes. The corpus totalizes 46689 words and was shared out among 3 categories: *Small*: less than 1000 words (15 texts); *Medium*: between 1000 and 3000 words (10 texts); and *Large*: more than 3000 words (5 texts). Each text has been annotated by 3 different annotators from a group of 5 with a random distribution between annotators in each size categories. Our 5

annotators were students in the master degrees of Linguistic and Computer Science.

The annotation was performed under the *Glozz* platform[1]. *Glozz* is based on a generic meta-model which allows to define any specific set of units (segments) and relations with editable features. It proposes a graphical environment and an SQL export, allowing annotations mining through standard database tools [8].

### B. Rhetorical and annotation model

The approach of rhetorical structure we consider is coarse-grained and segment-oriented (rather than relation-driven and bottom-up oriented as in discourse models such as RST [9] or SDRT [10]). Generally speaking, a *rhetorical segment* can be defined as filling a specific communicative function. Such segments can be defined in different ways.

One, following [2], is based on the notion of *discourse move*. Moves are conventional parts of the message, specific of a given genre[2]; for example, in scientific articles: context of the study, aim and hypotheses, experiments, results and discussion. In NLP, such a model has been notably worked out in [3] and adapted to other kinds of texts, such as administrative letters, in [4].

Another approach, in a sense more "universalist" is the classic *Narration-Description-Argumentation-Prescription* model [5], [6], [7]. Such *discourse modes* (according to Smith's denomination) may be considered as characterizations applying to full texts or, better, to parts of them. This model appeared to be well suited to our corpus and to the practical goals in view.

However, some adaptations were made. We observe that, in general, several discourse modes are simultaneously present in a same portion of text; for example description is intertwined with argumentation, or with narration. Rather than defining single characterizations of text segments (descriptive *or* argumentative *or* narrative...), discourse modes rather act as "colors" or "shades" that can combine.

Thus, the task of *rhetorical tagging* is described as follows. We make the hypothesis that paragraphs can be considered as relevant textual units: clearly, this hypothesis could be reconsidered but it seems an acceptable first approximation. Rhetorical tagging consists in identifying which discourse modes are present in a given paragraph and with which intensity. We proposed a set of seven discourse modes divided into two main dimensions, *representational* (or *ideational*) and *interpersonal*[3]. Annotators had to allocate a score to seven fields representing the intensity these seven discourse modes.

*a) Representational dimension:* It concerns the semantic content of the message, the representations construed by the reader. Four graded fields where proposed relative to four rhetoric modes.

*Description:* Indicates the weight of factual information in the paragraph.
*Argumentation-Explanation:* Represents to which extent the paragraph is about convincing or explaining something to the reader. We considered that the mechanisms of argumentation and explanation are the same, even if the goals are not.
*Chronology:* Indicates the weight of temporally marked information in the paragraph.
*Prospection:* Represents to which extent the paragraph projects the reader into the future.

*b) Interpersonal dimension:* It concerns the relation between the writer and the reader in the communicative process and includes three fields:
*Personal commitment:* Does the paragraph reflect the author's personal opinion or is it rather presented as objective ?
*Prescription:* To what extent the paragraph is about advising or instructing the reader to do something ?
*Polyphony:* Indicates the weight of directly or indirectly reported speech in the paragraph.
Each of the seven fields is given a score between 0 and 2. **0:** the discourse mode is absent or marginally present; **1:** it is present, but not essential to understand the paragraph; **2:** it constitutes a major key to understand the paragraph.

### C. Annotators agreement

We are in the standard situation of a known set of items (for each paragraph, the seven fields corresponding to the seven discourse modes) that receive some "tags" (0, 1 or 2 reflecting the presence and intensity of the mode in this particular paragraph) so that a Kappa measure will do well. *Fleiss' kappa* [11] coefficients for each text are presented in table I. For each text, the given score is the mean value of the scores of all its segments.

TABLE I
FLEISS *kappa* ON RHETORICAL MODES

| Text | $\kappa$ | Text | $\kappa$ |
|---|---|---|---|
| Large | | Small | |
| 0101_31 | 0.23 | 0101_1 | 0.42 |
| 0108_96 | **0.48** | 0101_14 | 0.41 |
| 0204_34 | 0.31 | 0101_20 | 0.36 |
| 0628_49 | 0.29 | 1229_33 | 0.39 |
| 1220_101 | 0.28 | 1230_90 | 0.33 |
| Average | 0.32 | 1231_75 | 0.49 |
| Medium | | 1231_86 | 0.32 |
| 0126_121 | 0.26 | 1231_93 | 0.39 |
| 0718_138 | −0.01 | 1231_70 | **0.59** |
| 0820_12 | **0.41** | 1231_84 | 0.52 |
| 0831_135 | 0.23 | 1231_89 | 0.30 |
| 1230_24 | 0.21 | 0101_13 | 0.37 |
| 0131_108 | 0.31 | 0101_19 | 0.36 |
| 0801_108 | 0.06 | 0101_6 | 0.23 |
| 0829_120 | 0.32 | 1230_3 | 0.28 |
| 0929_135 | 0.15 | Average | 0.39 |
| 1231_92 | 0.35 | | |
| Average | 0.23 | | |

The scores show an average agreement quality ranging from a "low fair" (medium texts) to "almost moderate" (small

[1]http://www.glozz.org/

[2]Where "genre" should be interpreted in a very narrow sense, and better called "micro-genre"

[3]The term "representational" is inspired by Adam's terminology and "ideational" by Halliday's one.

texts)[4]. They may look "modest", but one must keep in mind the highly interpretative nature of the task. Also remind that annotators had to choose between three possibilities; when considering only if the rhetorical color is present or absent all $\kappa$ raise by, at least, 0.1 (except for one text), leading to a global factor of 0.42, "moderate". The worst score on medium texts seems - according to the post-annotation debriefing - due to a particularly open interpretation of some of these texts.

On the overall, these results seem to show that a form of "convergence" does exists, pleading for the relevance of the model. Improvements could probably be obtained thanks to a better and less ambiguous definition of the discourse modes, for which a careful analysis of the present discrepancies should be helpful. Also differentiated analyses according to the different modes could be interesting: are some of them less controversial than others?

## III. The dynamic of discourse modes

All along this section we will use abbreviations indicated in table II to designate the rhetorical modes: DESC for Description, ARGU for Argumentation-Explanation, etc. The first column of the table presents a first bunch of raw observations, namely the distribution of modes along all texts and all annotators. Let us insist that it counts *scores*, i.e. the number of paragraphs annotated according to a particular mode, ponderated by the intensity factor given by the annotator. The sum of all scores for one mode is compared to the sum of all scores for all modes.

Two modes, DESC and ARGU are massively preeminent, while interpersonal ones are globally under-represented. This is clearly related to the kind of texts in the corpus, rather of "objective" nature, which do not include editorials for example, or forums. Let's consider now more elaborate questions.

### A. Preferred positions of discourse modes.

Figure 1 shows the repartition of rhetorical modes (evaluated as always in terms of scores) in the beginning, middle and final parts of texts. The global impression may corroborate primary intuitions. One can see two symmetric groups: CHRONO and DESC prefer the beginning with no marked preference for the other two: they present "facts" to be discussed later; while POLY, ARGU, PROS and PRES (which constitute such discussions) rather occur in the end. COMM is more equally distributed.

### B. Interdependence relations between discourse modes

Some correlations, positive or negative, between discourse modes may be conjectured from the previous table and observed on specific texts. For example a kind of contravariance between DESC and ARGU. The question arises whether such observation could be confirmed and generalized some way: are DESC and ARGU "generally" contravariant ? Are there other such pairs? In order to investigate this question we computed a statistical correlation coefficient.

[4]According to the interpretation grid of [12]: $\kappa < 0$: poor, $0 < \kappa < 0.2$: slight, $0.2 < \kappa < 0.4$: fair, $0.4 < \kappa < 0.6$: moderate, $0.6 < \kappa < 0.8$: substantial, $0.8 < \kappa < 1$: almost perfect

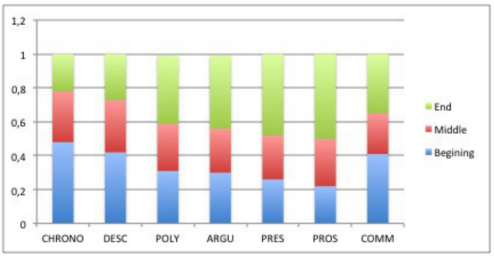


Fig. 1. Distribution of discourse modes along texts.

Results are shown in table III, limited to the four most representative rhetorical modes.

Three, low but perceptible correlations appear: a negative one between argumentation and description, as expected from our "manual" observations; a positive one between argumentation and personal commitment, which is not surprising; and a negative one again, between personal commitment and polyphony: stating other people positions is somewhat exclusive from expressing one's own. It is worth mentioning that, if the correlations are weak in value, the annotators agree on their direction, positive or negative, with one exception out of 18 pair annotator/mode: a fact that tends to strengthen the relevance of the results.

TABLE II
RHETORICAL MODES AND TOPIC TRANSITIONS. (1) THIS MODE/ALL MODES, ANYWHERE. (2) THIS MODE/ALL MODES, RESTRICTED TO TRANSITIONS. (3) THIS MODE IN TRANSITIONS / THIS MODE ANYWHERE. DESC: DESCRIPTION, ARGU: ARGUMENTATION-EXPLANATION, CHRO: CHRONOLOGY, PROS: PROSPECTIVE, POLY: POLYPHONY, PRES: PRESCRIPTIVE, COMM: PERSONAL COMMITMENT

| | (1) | (2) | (3) |
|---|---|---|---|
| DESC | 0.25 | 0.26 | 0.62 |
| ARGU | 0.27 | 0.28 | 0.6 |
| PROS | 0.06 | 0.04 | 0.41 |
| CHRO | 0.07 | 0.06 | 0.52 |
| POLY | 0.13 | 0.15 | 0.68 |
| PRES | 0.1 | 0.08 | 0.45 |
| COMM | 0.12 | 0.12 | 0.59 |
| ALL | 1 | 1 | 0.58 |

TABLE III
CORRELATION BETWEEN RHETORICAL MODES.

| | ARGU | DESC | COMM | POLY |
|---|---|---|---|---|
| ARGU | - | -0,18 | 0.22 | 0 |
| DESC | - | - | -0,08 | 0,08 |
| COMM | - | - | - | -0,16 |
| POLY | - | - | - | - |

### C. Rhetorical profiles.

One can produce graphics such as figures 2 and 3, useful to study the distribution of rhetorical modes along a given text. Rates of the 7 rhetorical modes (vertical axe) are displayed in relation with the successive paragraphs (horizontal axe)[5]. The first graphic concerns a historical-narrative text: the biography

[5]In this example, each graphic concerns a single annotator.

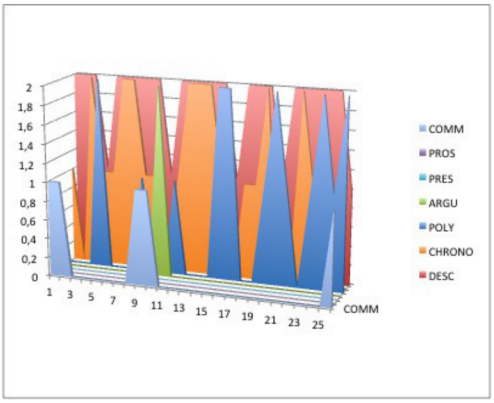


Fig. 2. The dynamic of discourse modes: Biography of an Israel spy.

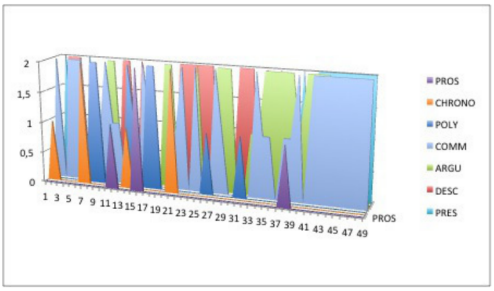


Fig. 3. The dynamic of discourse modes: "How to save Bosnia".

of an Israel spy; the second is an analytic paper about the war in Bosnia. The first thing that appears is the difference in their dominant modes: DESC, CHRONO and POLY for the first, PRES, DESC and ARGU for the second, with also COMM, which more or less coincides with PRES and is hidden by it on the figure. A closer look shows informations about the plan of the texts.

In figure 2 we have a "ground" of descriptive-chronological discourse all along the text, with mainly in the second half a strong polyphonic component (which correspond to discussions and conjectures about the "real" life and activity of this spy). In figure 3 we have a concentration of prescriptive and argumentative discourse (with strong personal commitment) in the second part, while descriptive-chronology-polyphony is rather concentrated in the beginning (stating the history and the problem). These quick observations tend to show, first that a rhetorical dynamic can be analyzed in terms of discourse modes, and second that "text profiles" can be inferred.

## IV. Discourse modes and text segmentation

Going further, we can consider the question whether discourse modes by themselves may determine text segmentation i.e. allow to define a succession of segments being "rhetorically homogeneous" in some way. This question was addressed in two different assumptions: a strong one, where a segmentation would be (strictly) determined by changes in the configuration of salient discourse modes - the result seems to be negative; and a weak one, where we consider the contribution of discourse modes to some general "thematic segmentation" - and interesting correlations can be put to light.

### A. Attempt towards a pure rhetorical segmentation

We imagined the following experiment. Drawing a parallel with conventional methods in thematic segmentation [1] we considered rhetorical modes as a set of descriptors: the scores given by an annotator to some paragraph defines a *vector* which represents its rhetorical orientation. We can compute angles between successive blocks and deduce continuity or discontinuity according to the angle being smaller or greater than some threshold. Unfortunately the first results are not very convincing: if some "relatively homogeneous" contiguous regions of some extent (several paragraphs) may appear, such text ranges are rather scarce. More subtle measures could probably be considered but in fact the transcription of topical segmentation techniques, based on a geometry of rhetorical descriptors, does not seem to be the good idea.

### B. Topical structures and discourse modes

Another way to consider the role of discourse modes in text organization is to look for possible correlations with the topical structure of a text. Here, we took advantage of another annotation of the corpus, performed simultaneously, where annotators were asked to cut out the texts into geat "parts" according to their reader's intuition [13] . This could be called "spontaneous segmentation", as performed by an attentive reader, more or less consciously. The general question is to know what, in this operation, is relative to the "subject" (knowledge domain, discourse referents...) and what to rhetorical features. We were in this matter especially interested in the transitions between "spontaneous" segments and asked the annotators to mark sentences that, according to them, signaled such transitions[6].

Two kinds of investigations were made. The first one continues the geometrical model as above ($A$), searching for a possible coincidence between rhetorical gaps - measured as angles between "rhetorical vectors" - and topical changes. Unfortunately, at first sight at least, the test fails: there are cases where topical changes are accompanied with large rhetorical angles and cases where not. And the core of topical segments shows both low and hight rhetorical gaps. Again, looking for global configurations of discourse modes appears not to be the right way.

Then we decided to have a more individuated view on each discourse mode and to concentrate on annotated *transitions*: do such zones have specific rhetorical characteristics? Results are figured out in columns 2 and 3 of table II ("Transition" means "paragraph containing an annotated transition zone"). Column $(2)$, when compared to $(1)$, would induce a rather negative result: modes distribution does not reveal significant difference between transitions and other blocks. But (3) shows more positive results.

[6]See for instance [14], [15], [16] for studies on the linguistic characteristics of topical changes.

First we see that introductory blocks contribute for 0.58 to the total score: an important ratio since they constitute only one third of all paragraphs. Beginnings of topical segments are more strongly marked than the others on the rhetorical ground. Then we see that POLY and, with lower strength, DESC clearly prefer transitions. On the opposite side, PROS, PRES and to some extent CHRO are less represented in this position. In other words writers like to begin a new topic by the presentation of different viewpoints or descriptive information and tend to reject prospective, prescriptive or chronological considerations out of this position.

Hence, as one could expect and despite the results of our first test, there seems to be real hints for an implication of rhetorical concerns in topical organisation. On a practical ground, these results could also help in automated segmentation procedures, provided one could find reliable marks of the four distinguished modes, a question addressed in the last part of the paper.

### C. *Conclusion: what kind of* rhetoric zoning*?*

Gathering the results of sections 3 and 4, some information can be synthesized concerning the organization of discourse modes along texts. The negative - but, still, informative - result is that no clear segmentation (in contiguous blocks) is likely to be based on *global* configurations of discourse modes.

Contrastively, different experiments have shown that, taken *individually*, discourse modes do determine some *zones* according to their salience - which is most important w.r.t. our targeted application. And finally we have seen that the combination of topical and rhetorical features is relevant to the spontaneous segmentation of texts by readers, which might provide hints for improvements in thematic segmentation.

## V. Towards an automatic tagging of discourse modes

### A. *Procedure and implementation*

An automatic tagging of text passages w.r.t. the given set of discourse modes appears as a necessary complement to the previous study, both in order to contribute to the validation of the model, and as the basis for the application in view. A first step was performed in this direction as follows [17]. We listed a set of features whose count allows to assign a score to each mode in each paragraph. This score is supposed to reflect the force of the considered mode. In this first experimental attempt, we considered simple features, essentially lexical and morphological ones, as illustrated by the following sample.

- Description: verb tenses representing durative processes (imparfait, présent[7]), spatial locative connectives (prepositions sur/ *on*, dans / *in*...), adjectives and relative pronouns, demonstrative pronouns, named entities.
- Chronology: verb tenses of the past (passé simple, passé composé, participe passé), temporal connectives (conjonctions and adverbials: quand / *when*, puis / *then*, ce matin / *this morning*...), dates.
- Prospection: verb tenses of future and unrealised (futur, conditionnel), cue words (à l'avenir / *in the future*, hypothèse / *hypothesis*, prévoir / *forsee*...)
- Argumentation-Explanation: logical and argumentative connectives (cependant / *however*, donc / *hence*, d'abord / *first*, ensuite / *then*, parce que / *because*...), other cue words (impliquer / *imply*, problème / *problem*, réponse / *answer*...)
- Polyphony: quotation marks, proper names and social functions (as indicating possible authors of reported speech), declarative verbs.
- Prescription: verb mode (impératif), modal verbs (pouvoir / *can*, falloir / *must*, devoir / *must*), other cue terms (important / *important*, essentiel / *essential*...)
- Personal Commitment: logical connectives, epistemic verbs at 1st person (penser / *think*, douter / *doubt*...), other cue terms (respect / *respect*, inquiétude / *concern*...).

One can remark that some features are shared by several modes: it is the co-presence of other ones of the same family that determines *in fine* their rhetorical orientation. The scoring takes this phenomenon into account as illustrated by the following example.

**Text :** En août, explique Hugues Portelli, qui veille sur les courbes d'opinion au sein du cabinet du premier ministre, il y a eu le consensus monétaire après la crise de juillet et, pour cette fin d'année, le premier ministre a profité, à la fois, de la gestion du conflit d'Air France, des actions menées par Charles Pasqua, notamment contre le FIS, et, enfin, du résultat obtenu sur le GATT.

*In August, says Hugues Portelli, who watches over the curves of opinion within the PMO, there was consensus after the monetary crisis in July and, for this season, the Premier took the opportunity in the same time of the conflict management at Air France, the actions taken by Charles Pasqua, especially against the FIS, and finally the result of the GATT.*

- Clues for Polyphony: Hugues Portelli, Charles Pasqua [named entities for persons], Expliquer [cue verb]. Score = 3.
- Description: Hugues Portelli, Charles Pasqua, Air France, FIS, GATT [named entities], qui [relative pronoun], sur, au sein de [spatial connective], premier ministre [function name], et [conjonction] (twice), Score = 11.
- Argumentation-Explanation: Expliquer [cue verb], après, et (twice), pour, à la fois, notamment, enfin [logical connectives]. Score = 8.
- Chronology: il y a eu, profité, obtenu [past tense], en août , après, juillet, fin, année, à la fois, enfin [temporal terms]. Score = 10.
- Other modes: 0 clue.

The procedure was tuned on a corpus made of some thirty texts from *Le Monde*, disjoint from the annotated ones, then tested on the later. In this first experiment we limited ourselves to identify the three modes most representative of each paragraph (Description, Argumentation-Explanation and Chronology in the example). An XML output allows to insert

[7]For obvious reasons, tenses are given their French name.

TABLE IV
AUTOMATIC AND HUMAN RHETORICAL TAGGING OF A TEXT

| | Annot.1 | Annot. 2 | Annot. 3 | *Automated Annot.* |
|---|---|---|---|---|
| §1 | 1. ARGU<br>2. DESC<br>3. POLY | 1. ARGU<br>2. COMM<br>3. POLY | 1. POLY<br>2. PRES | *1. ARGU*<br>*2. DESC*<br>*3. POLY* |
| §2 | 1. ARGU<br>2. DESC<br>3. COMM | 1. ARGU<br>2. COMM<br>3. POLY | 1. PRES<br>2. ARGU | *1. ARGU*<br>*2. DESC*<br>*3. POLY* |
| §3 | 1. ARGU<br>2. COMM | 1. COMM<br>2. PRES<br>3. ARGU | 1. PRES<br>2. ARGU | *1. DESC*<br>*2. ARGU*<br>*3. COMM* |

this result in the text.

### B. First results and evaluation

We compared our automatic labeling to the manual annotations of the corpus. The result, still qualitative, is that our annotation does not scatter more than what can be noted between human annotators themselves (see an example for a single text in Table IV). In other words, the automatic tagging does not seem better or worse that the manual ones, and is in fact consistent with them. If the evaluation procedure clearly requires to be refined, as well as the manual tagging itself, we believe that this first test may be considered encouraging concerning the feasibility of the task. An important issue to highlight is the underrepresentation of certain modes: Prospection and modes of the interpersonal dimension (with the exception of Polyphony). In the future we therefore need to go beyond the three prevailing modes and probably separate representational and interpersonal ones.

## VI. CONCLUSION AND FUTURE WORK

In this paper we have presented a set of investigations on the rhetorical structure of journalistic texts, based on the constitution of an annotated corpus Our first positive result consists in the corpus itself since there is a recognized lack for such resources[8]. The inter-annotators agreement seems acceptable, considering the strongly interpretative nature of the task: generally speaking we believe that, in the case of discourse structure, we have to learn how to cope with this variability rather that try to reduce it to null.

The rhetorical model was designed in terms of discourse modes, due to the application in view, a specific kind of document browsing adapted to journalistic texts. It was correctly received by the annotators, which provides an encouraging hint of its relevance. In particular the idea of a combination of discourse modes in a same passage, with graduation of their salience, seems to receive confirmation.

Several quantified observations, performed thanks to the annotated corpus, give useful informations on the distribution of discourse modes and their contribution to text zoning of some kind. A notion of "rhetorical profile" emerges, combining global dominant modes with the "dynamic" of modes distribution along the text.

[8]Glozz annotations are "stand off" and may be obtained for free from the authors, provided the applicant has him/herself acquired the rights on *Le Monde* corpus.

Finally, a first step was performed towards an automatic tagging in relation to the model.

Further work should include the following questions.

**1.** An extension of the corpus, in order to give a firmer value to our analyses. A careful examination of the discrepancies between annotators could provide useful hints in order to tune the model and remove ambiguities in its description. Achieving a better inter-annotators agreements would be a good confirmation of these improvements. Another issue would be to split the corpus into different more homogeneous subtypes.

**2.** Improving the automatic labeling. Other linguistic parameters should be considered. Especially interesting would be aspectual values as described in [6]. Machine learning issues should be considered, but need clearly a great effort in corpus annotation.

**3.** Finally, the application itself should be considered, which implies to convert "pragmatic" requirements of readers into configurations of rhetorical modes.

## REFERENCES


[1] M. A. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages." *Computational Linguistics.*, vol. 23, no. 1, pp. 33–64, 1997.

[2] J. Swales, *Research Genres: Exploration and Application*. Cambridge University Press, 2004.

[3] S. Teufel and M. Moens, "Discourse-level argumentation in scientific articles: human and automatic annotation." *In Proceedings of ACL-99 Workshop "Towards Standards and Tools for Discourse Tagging"*, pp. 84–93, 1999.

[4] D. Biber, U. Connot, and T. A. Upton, *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. John Benjamins Publishing Co, 2007.

[5] E. Werlich, *Typologie der texte*. Quelle and Meyer, 1975.

[6] C. S. Smith, "Discourse modes: aspectual entities and tense interpretation," *Cahiers de Grammaire*, vol. 26, pp. 183–206, 2001.

[7] J. M. Adam, *La linguistique textuelle. Introduction à l'analyse textuelle des discours*. Armand Colin, 2005.

[8] A. Widlöcher and Y. Mathet, "La plate-forme glozz: environnement d'annotation et d'exploration de corpus," *Actes de TALN'09*, 2009.

[9] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text*, vol. 8, no. 3, pp. 243–281, 1988.

[10] N. Asher, *Reference to Abstract Objects in Discourse: A Philosophical Semantics for Natural Language Metaphysics*. Kluwer Academic Publishers, 1993.

[11] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[12] J. Landis and G. Koch, "A one-way components of variance model for categorical data," *Biometrics*, vol. 33, pp. 671–679, 1977.

[13] A. Labadié, P. Enjalbert, and S. Ferrari, "Transitions thématiques : Annotation d'un corpus journalistique et premières analyses (manual thematic annotation of a journalistic corpus : first observations and evaluation) [in french]," in *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN*. Grenoble, France: ATALA/AFCP, June 2012, pp. 503–510. [Online]. Available: http://www.aclweb.org/anthology/F/F12/F12-2046

[14] M. Charolles, "L'encadrement du discours : univers, champs, domaines et espaces." *Cahier de Recherche Linguistique*, vol. 6, pp. 1–73, 1997.

[15] N. Asher, P. Dennis, and B. Reese, "Names and pops and discourse structure," in *Workshop on Constraints in Discourse*, Maynooth, July 2006, pp. 11–18.

[16] S. Piérard and Y. Bestgen, "Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de textes," *TAL*, vol. 2, no. 47, pp. 89–110, 2006.

[17] A. Attoumani, "étiquetage d'un texte selon différents modes rhétoriques. master's thesis." University of Caen, Tech. Rep., 2011.