Exploitation of speculation markers to identify the structure of biomedical scientific writing

Natalia Grabar[1,2], PhD, Thierry Hamon[3], PhD

[1]**Centre de Recherche des Cordeliers, Université Pierre et Marie Curie - Paris6, UMR_S 872, Paris, F-75006; Université Paris Descartes, UMR_S 872, Paris, F-75006; INSERM, U872, Paris, F-75006 France**

[2]**HEGP AP-HP, 20 rue Leblanc, Paris, F-75015 France**

[3]**LIM&BIO (EA3969), Université Paris 13, 74, rue Marcel Cachin, 93017 Bobigny Cedex France**

Abstract

The motivation of this work is to study the use of speculation markers within scientific writing: this may be useful for discovering whether these markers are regularly spread across biomedical articles and then for establishing the logical structure of articles. To achieve these objectives, we compute associations between article sections and speculation markers. We use machine learning algorithms to show that there are strong and interesting associations between speculation markers and article structure. For instance, strong markers, which strongly influence the presentation of knowledge, are specific to RESULTS, DISCUSSION *and* ABSTRACT*; while non strong markers appear with higher regularity within* MATERIAL AND METHODS. *Our results indicate that speculation is governed by observable usage rules within scientific articles and can help their structuring.*

Introduction

Structure of scientific writing is helpful for the information extraction.[1] However, it is not always available: document structure may be removed during processing, *e.g.* conversion from `pdf` or `word` files. In order to recover the structural information (*i.e.*, hypotheses, results, perspectives, etc.), we propose to rely on speculation markers, or linguistic hedges,[2] which are widely used by scientists in a variety of contexts:

1. Thus, eIF5 and Sui1p may be recruited to the 40S ribosomes through physical interactions with the Nip1p subunit of eIF3.
2. RGS4 is palmitoylated, with Cys-2 and Cys-12 the likely sites of palmitoylation.
3. A model for the regulation of hypoxia-inducible mRNAs by pVHL is presented based on the apparent similarity of elongin C and Cul2 to Skp1 and Cdc53, respectively.
4. Based on these results, we propose that Nipped-A, through the action of the SAGA and Tip60 complexes, facilitates assembly of the Notch activator complex and target gene transcription.
5. However, although we favor an initial role for eIF4E-T and/or rck/p54, the order of mRNA remodeling events is presently not clear.

It has been proposed that speculation speech is used for marking rhetorical developments and moves within the scientific literature,[3] for making categorical assertions softer,[2] or for ensuring coded reading of the literature.[4] There is a growing interest in the study of the speculative markers from both descriptive and application points of view. The description of hedges focuses on their linguistic and semantic characteristics. A study of part of speech (POS) categories shows that hedges are not category-specific, but are distributed among various POS categories[2]: lexical verbs (*seem, suggest, appear*), adverbs (*certainly, possibly, likely*), adjectives (*possible, plausible, putative*), modal verbs (*can, might, should*) and nouns (*assumption, hypothesis, suggestion*). The study of semantics has been addressed in a variety of ways. Hedges have been classified according to the rhetorical zones they appear in within the scientific literature and according to the level of confidence conveyed by the markers. For instance, the "zone analysis"[5] distinguished classes of markers likely to signify *Problem setting*, *Insights*, *Background* or *Implication*. Objective of such a work is to pinpoint and organize factual information (such as experimental results) by enriching information extraction systems with rhetorical zones and thus by classifying spans of text in terms of argumentation and intellectual attribution. As for the graduation of hedge confidence, it consists of associating marker with the degree of confidence they usually convey when used with a given information. For instance, among modal markers,[6,7] *must* is found to convey stronger confidence than *should*, *should* than *may*, *may* than *could* or *might*, etc. The classification is more difficult to tackle within a larger set of markers, where various types of characterization should be implemented, such as *knowledge type*, *certainty level* or *point of view*.[8] On a sample of 202 abstracts, manually annotated in the last work, the authors noticed that the distribution of markers is not equal: 77% belong to *knowledge type*, 16% to *certainty level*, and 7% to *point of view*. Within the *certainty level* markers, authors distinguished absolute, high, moderate and low degrees of certainty. For instance, *probable* conveys more confidence than *possibly* or *unlikely*. It is obvious that semantic characterization of speculation markers

is dependent on objectives, points of view, applications, etc. From an application perspective, speculation markers have been often used for the detection of specific pieces of information: speculative sentences,[9] citation contexts,[10] sentences for the automatic generation of literature abstracts,[11,12] characterization of gene annotations,[1] and rhetorical moves.[5,3] In spite of being difficult to classify and manage within documents, markers may provide rich and meaningful insights into scientific writing.

In this paper, we explore a novel aspect of speculation markers used in scientific writing. Specifically, we study the relation between speculation markers and the discursive structure of articles. This aspect has not been studied in the state of the art. In addition, our work relied on full text articles, whereas previous studies mainly focused on abstracts.

Material and Methods

For detecting the relation between speculation markers and the discursive structure of articles, we propose to categorize article sections based on the speculation markers they contain and to compare an automatically obtained categorization with the reference (articles structured by their authors). The methods we propose rely on the following elements: automatic categorization, corpus, categorization features, and evaluation.

Automatic categorization. We apply several classification algorithms for supervised machine learning as they are implemented within the `Weka` platform.[13] Learning algorithms consider documents (article sections) as vectors within a vector-space. The dimension of this space depends on the number of vectors or features (here, speculation markers). The size of each vector corresponds to the frequency of a given marker in a given section. The main challenge of the method lies in data preparation: (1) correct recognition of various sections within scientific articles, and (2) selection and classification of speculation markers.

Corpus: article sections. In this study, we used a corpus of the literature related to genes and gene products associated to neurodegenerative diseases in humans, such as Alzheimer's disease. Articles in the corpus were collected through PubMed,[14] thanks to gene clusters.[15] A total of 355 PubMed citations were obtained, including 41 that are available as full text HTML documents through the Pubmed Central.[16] `Perl` scripts were used to collect these articles and to segment them into sections relying on the HTML-tree structure and section names. The study focused on sections that appear the most regularly: ABSTRACT, INTRODUCTION, MATERIAL AND METHODS, RESULTS and DISCUSSION. Other sections (ACKNOWLEDGEMENT, REFERENCES, TABLES, FIGURES or NOTES) are neither regular nor frequent enough to be processed with the same method. The studied articles were published between 1987 and 2006 in 14 journals. In spite of the relatively small size of the analyzed corpus, we expect it provides a representative sample of a variety of idiolects and styles used by researchers and journals in experimental biology.

Features: speculation markers. We use a set of 363 speculation markers manually collected from biomedical articles.[7] These markers belong to various POS categories (verbs, nouns and noun groups, adverbs, adjectives, modal verbs, punctuation, adverbial and prepositional groups). They can be simple words or complex expressions. In our work, we distinguish three types of markers according to their influence on knowledge:

1. strong markers that strongly modify confidence of the information (*i.e.*, *assumption, can, may, putative, presumed*);
2. weak markers that convey little influence on information although their presence can be meaningful (*i.e.*, *almost, although, believed to, clear, despite, however, relative, surprising*);
3. intermediate markers are all the remaining markers, which are not easy to assign to the two previous types (*i.e.*, *appear, clues, deduced, expect, further analysis, future, here, induced*).

Based on this typology, we use three sets of markers: *all* (n=363), *strong* (n=59) and *non strong* (n=304) markers. The last set (*non strong* markers) includes both intermediate and weak markers. These three sets (*all*, *strong* and *non strong*) are exploited with or without the size of segments s processed (number of occurrences or words they contain).

Evaluation. Results obtained with machine learning algorithms are evaluated according to the reference data, *viz.* the original structure of scientific articles. Training and testing are performed on independent corpora, using 10-fold cross-validation. The evaluation of categorization results is performed with two classical measures: precision (percentage of correct items among all the items categorized) and recall (percentage of categorized items among the items that should be categorized). We mainly used macro precision and recall scores: mean values are computed for each category aimed (sections of articles). However, when relevant, micro precision and recall are also given: they correspond to the mean values obtained at the level of article sections (and not categories).

Results and Discussion

Among the learning algorithms tested, Classification Via `Regression` provides the best results. Graphs for each set of features exploited (*all*, *strong* and *non strong*, with or without the text size s of segments) are presented on figure 1, along with mean values of macro-recall and macro-precision. Categorization is addressed here as a multicategorization problem: one model is generated for distinction of all five types of sections. Results obtained with features only are shown in the left column (fig. 1(a), 1(c), 1(e)), results obtained with features and the text size s are shown in the right column (fig. 1(b), 1(d), 1(f)). The various sections are abbreviated as follows: *AB* (ABSTRACT), *IN* (INTRODUCTION), *MM* (MATERIAL AND METHODS), *RS* (RESULTS), *DS* (DISCUSSION). Results are given in terms of macro precision and recall. In these graphs, the better the categorization results (both precision and recall), the closer they are to the top-right corner. As for the relation between speculation markers and sections, when a section is correctly recognized it means that speculation markers are specific to this section; in other words, it means that there is a strong association between them and that the approach can be used for structuring purposes. In other cases, the association between sections and markers used is weak and inconclusive.

When all 363 features are used (fig. 1(a)), the best association between sections and speculation markers is found within *MM* (R=0.975, P=0.907) and *AB* (R=0.977, P=0.754). These are followed closely by *RS*, and then by *DS*. The weakest association is shown within *IN*. This means that the total set of speculation markers is used with the best regularity within *MM*, *AB* and *RS*. When we make a distinction between *strong* (fig. 1(c)) and *non strong* (fig. 1(e)) markers, we observe that *strong* markers are strongly associated with *RS* (R=0.750, P=0.732) and *DS* (R=0.634, P=0.743), and slightly with *AB*. They are not at all associated with *MM* or *IN*. As for *non strong* markers, their best association is observed with the *MM* (R=0.950, P=0.927). In all these experiments, *IN* makes either little or irregular use of speculation markers. When, in addition to speculation markers, we take into account the text size, categorization results improve by 0.05 to 0.10 for recall and up to 0.50 for precision (fig. 1(b), 1(d), and 1(f)). Text size gives additional information on the nature of text segments. Its main contribution is that it allows to correctly distinguish *AB*, which is always a short piece of text. The best categorization results are obtained with all 363 features and text size, *i.e.* when the most extensive learning information is available. In this case, the recognition of *MM* and *AB* is very close to the reference data.

A large number of features is used (n=363) especially by comparison with the number of processed documents (n=41) or the number of processed segments (n=205). However, only a small number of features is actually used by the categorization algorithms for the creation of learning models: *not, can, also, shown* and *finding* among *all* features, *whether, could, may* and *should* among *strong* features, and *also, described, shown* and *but* among *non strong* features. Text size is always significant. The features selected by the algorithms may reflect the most regular patterns of hedges within the articles processed. We believe they would be valid for the processing of a larger set of articles.

We observed that *strong* markers are mainly specific to *RS*, *DS* and *AB*, and *non strong* markers to *MM*. Thus, *strong* markers seem to be used for the presentation and discussion of new experimental results: authors appear to be particularly careful when presenting this type of information. Although *RS* and *DS* sections were found to make a similar use of markers of the *strong* type, the markers themselves are different. As a matter of fact, we observe a greater variety of markers within *DS* sections and, among them, we find more conditional markers (*may*, *could*, *would*) and verbs like *imply* or *speculate*. We also noticed that the text size is larger for *RS* than for *DS*. Besides, decision features for *RS* are more discriminatory and higher in the decision tree than those used for *DS*. As for *AB* sections, they are usually ambiguous with different sections, and especially with *RS* and *DS*. Indeed, abstracts of scientific papers represent the structure and the main content of the whole paper; and this can be easily observed through the use of speculation markers. Finally, *MM* of biological articles focus on the presentation and elaboration of methods and their relation with other work. According to our results, the use of speculation markers within scientific articles in biology is not arbitrary and appears to be guided by rhetorical and stylistic rules. From this point of view, speculation markers have a special role: they give rhythm to scientific writing, emphasize precise information and may contribute to the detection of logical structure of articles.[17]

Conclusion and Perspectives

We addressed the regularity of use of speculation markers within full text articles, as it can be observed through automatic categorization. Our experiments showed that speculation markers are widely and regularly spread across scientific articles in biology, and can indeed be used to categorize article sections. *Strong* markers appear to be specific to RESULTS, DISCUSSION and ABSTRACT sections; while *non strong*

markers are specific to MATERIAL AND METHODS.

A document is a complex entity. Finding statistical patterns, such as those described in this work, is a first step in understanding document structure and content. Our study has been performed with segments corresponding to entire sections which allowed to observe the specificity of speculation markers at this macro-level. From the point of view of structuring of articles, a similar experiment can be performed at the level of paragraphs: it can be helpful for the structuring of articles when their text is extracted from the `pdf` or `word` files and loses its explicit logical structure.[17] This is the current perspective of our work. Finally, at the level of sentences, the objective would aim at the generation of abstracts[11,12] or extraction of pieces of knowledge and their characterization.[1,7,18] This perspective may help the information extraction process.

We used a set of speculation markers manually extracted from biomedical articles in our previous work,[7] but other sources[2,19] of markers can also be used. A comparison between markers we find within scientific writing and those already proposed by the state of the art can be performed. As we noticed in the description of material, we distinguished three types of speculation markers: *strong* (which strongly influence the knowledge), *weak* (which slightly influence the knowledge) and *intermediate* (which have an intermediate influence on knowledge). Markers of this last type are not easy to assign to any of the two previous types. During the experiments, we aggregated *weak* and *intermediate* markers within the same feature set. But it is obvious that we can use these two types separately, which would lead to a better understanding of markers and of their use.

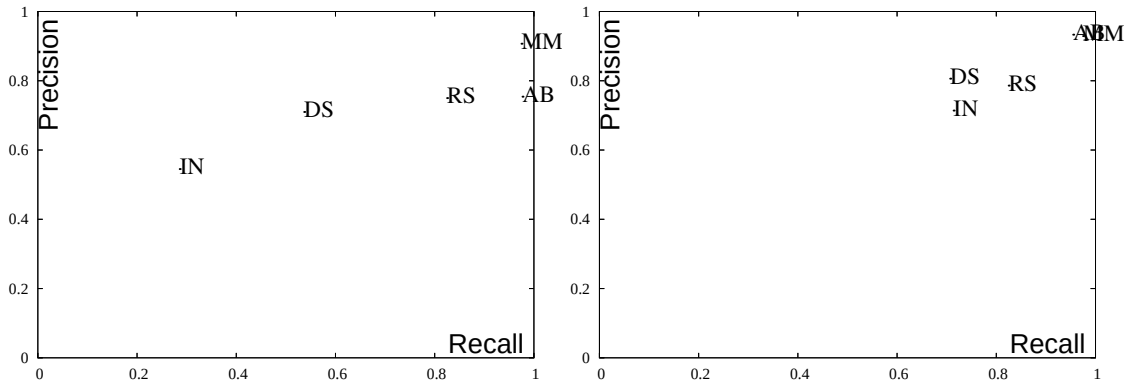
This study should be applied to a larger set of data and to articles from other areas (computational sciences, law, ...) as well as to clinical documents.[20]

Acknowledgments

We thank the anonymous reviewers for their helpful comments and Aurélie Névéol for her editorial assistance and further comments on the manuscript.

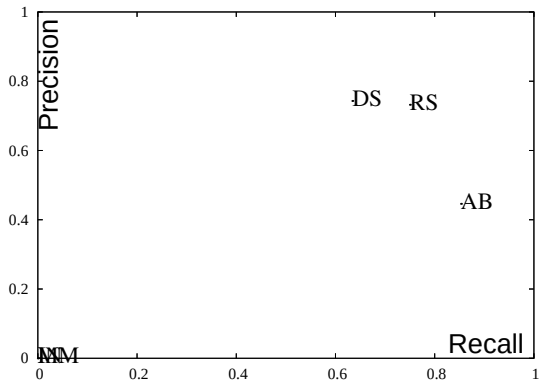
REFERENCES

1. Ruch P, Boyer C, Chichester C, et al. Using argumentation to extract key sentences from biomedical abstracts. *Int J Med Inform* 2006;76(2-3):195–200.
2. Hyland K. The author in the text : Hedging scientific writing. *Hong Kong papers in linguistics and language teaching* 1995;18:33–42.
3. Teufel S. Meta-discourse markers and problem-structuring in scientific articles. In: ACL Workshop on Discourse Structure and Discourse Markers, Montreal. 1998:43–9.
4. Fahnestock J. *Rhetorical figures in science*. Oxford University Press, 1999.
5. Mizuta Y, Korhonen A, Mullen T, and Collier N. Zone analysis in biology articles as a basis for information extraction. *IJMI* 2006;75:468–87.
6. Jacobs I. English modal verbs. Technical report, W3C, 1995. Available at www.w3.org/People/Jacobs/modals.ps.
7. Jilani I, Grabar N, Meneton P, and Jaulent MC. Classification of biomedical knowledge according to confidence criteria. In: Stud Health Technol Inform, 2008. To appear.
8. Thompson P, Venturi G, McNaught J, Montemagni S, and Ananiadou S. Categorising modality in biomedical texts. In: LREC workshop "Building and Evaluating resources for biomedical text mining", 2008.
9. Light M, Qiu XY, and Srinivasan P. The language of bioscience: facts, speculations and statements in between. In: ACL WS on Linking biological literature, ontologies and databases, 2004:17–24.
10. Mercer RE, Marco CD, and Kroon FW. The frequency of hedging cues in citation contexts in scientific writing. In: in Computer Science LN, ed, CSCSI. Springer Berlin, 2004:75–88.
11. Paice CD. The automatic generation of literature abstracts : an approach based on the identification of self-indicating phrases. In: SIGIR, Cambridge. 1980:172–91.
12. Teufel S and Moens M. Sentence extraction and rhetorical classification for flexible abstracts. In: Spring AAAI Symposium on Intelligent Text summarization, Menlo Park, CA. 1998:89–97.
13. Witten I and Frank E. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2005.
14. NLM . Medline: medical literature on-line. National Library of Medicine, Bethesda, Maryland, 2008. www.ncbi.nlm.nih.gov/sites/entrez.
15. Lefebvre C, Aude JC, Clément E, and Néri C. Balancing protein similarity and gene co-expression reveals new links between genetic conservation and developmental diversity in invertebrates. *Bioinformatics* 2005;21(8):1550–8.
16. NLM . Pubmed Central: medical literature on-line. National Library of Medicine, Bethesda, Maryland, 2000. www.pubmedcentral.nih.gov.
17. Adam C and Morlane-Hondère F. Détection de la cohésion lexicale par similarité distributionnelle : application à la segmentation thématique. In: RECITAL 2009, 2009.
18. Lu Z, Cohen K, and Hunter L. Finding generifs via gene ontology annotations. In: Pac Symp Biocomput, 2006:52–63.
19. Ruppenhofer J, Ellsworth M, Petruck MRL, Johnson CR, and Scheffczyk J. Framenet ii: Extended theory and practice. Technical report, FrameNet, 2006. Available online http://framenet.icsi.berkeley.edu.
20. Friedman C, Alderson PO, Austin JH, Cimino JJ, and Johnson SB. A general natural-language text processor for clinical radioology. *JAMIA* 1994;1(2):161–74.

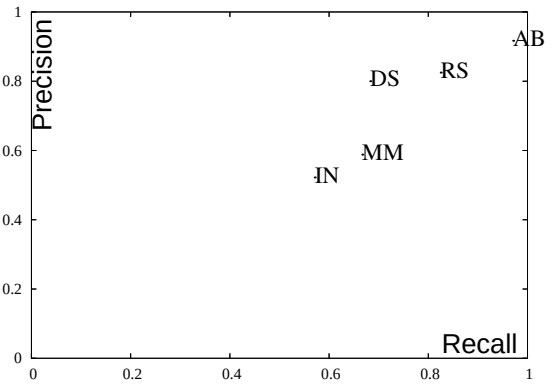


(a) All markers (maR=0.720, maP=0.733)

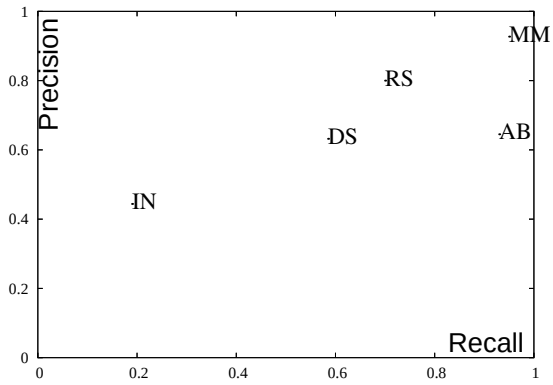
(b) All markers + size (maR=0.835, maP=0.833)



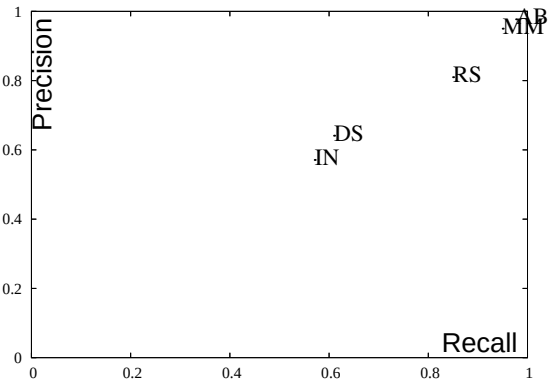
(c) Strong markers (maR=0.312, maP=0.385)



(d) Strong markers + size (maR=0.743, maP=0.730)



(e) Non strong markers (maR=0.671, maP=0.670)



(f) Non strong markers + size (maR=0.791, maP=0.789)

Figure 1: Categorization results obtained with the `Regression` algorithm. Each graph represents one of the six sets of features used: (a) *all*, (b) *all+size*, (c) *strong*, (d) *strong+size*, (e) *non strong*, (f) *non strong+size*). For each set of features used, we indicate mean values of macro-recall (maR) and macro-precision (maP).