"*Confortation*": about a new qualitative category for analyzing biomedical texts

Delphine Battistelli[1], Antonietta Folino[2], Patricia Geretto[2], Ludivine Kuznik[2], Jean-Luc Minel[1], Florence Amardeilh[1]

[1] MoDyCo-UMR 7114, Université Paris X -CNRS, F92001 Nanterre, France
delphine.battistelli@paris-sorbonne.fr, jean-luc.minel@u-paris10.fr, florence.amardeilh@mondeca.com
[2] Université Paris X, F92001 Nanterre, France
antonietta.folino@unical.it, pgeretto@hotmail.fr, ludivine.kuznik@gmail.com

Abstract: In this paper we present a new approach to the expression of certainty and uncertainty in scientific experimental articles. This will permit to ascertain the validity of knowledge extracted from biological literature and used to automatically populate a domain ontology. We argue that lexical terms such as *show, find, observe*… express a semantic category different from the one characterized by markers such as *demonstrate*, *validate*, *support*… We name the latter category *"confortation"* as it conveys a notion of strengthening and we propose five other semantic categories: lack of knowledge, objects of study, hypothesis, observations, and general knowledge. This last category and the linguistic phenomenon of reported speech are respectively examined as consensual truth and as knowledge reported from identified scientific sources.

Key words: epistemic modality, semantic annotation, certainty concept, uncertainty concept, biomedical articles, scientific discourse, *"confortation"*

1 Introduction and Context

This article has been carried out as part of the Microbio Project[1] which stands at the intersection of three areas of automatic natural language processing (Text Mining, Knowledge Acquisition, and Information Extraction) applied to biomedical research on miRNAs[2], molecules in which the biological community shows increasing interest because of their impact on the development or inhibition of certain diseases.

Biomedical research is progressing rapidly and involves more and more laboratories, leading to a dramatic increase in the amount of published information.

[1] http://www.microbioamsud.net/
[2] miRNAs (MicroRNAs): small RNA molecules encoded in the genomes of plants and animals

Biologists can no longer keep abreast of developments in miRNA research, making it impossible for them to identify and monitor information.

The difficulty in finding relevant information because of the large amount of available articles is also described in [1]. There the role of semantic annotation is to highlight key concepts and to facilitate skimming documents, in order to evaluate whether it is worth reading them fully.

In our context, the aim of semantic annotation consists in designing a tool to support biologists in their daily tasks for finding relevant textual parts of scientific articles. For that reason a collaboration between biologists (from the Pasteur Institute in Montevideo) and computer scientists and linguists (from MoDyCo) has been set up. An initial domain ontology about miRNAs was modeled by taking as input various existing Web resources (such as the databases, the Sequence Ontology…) and a set of expert interviews with the biologists. An information extraction tool aiming at automatically populating this miRNA ontology was configured to identify and semantically annotate sentences which contain the name of a miRNA and at least a Gene or a Mutation [2]. The biologists pointed out however that this tool was insufficient to ascertain the validity of the new knowledge

One difficulty in the search[3] for useful knowledge is that information is scattered throughout the text. In fact, as explained in [3], temporal, modal and/or enunciative features indicative of authorial commitment to the identified information in biological text need to be explicitly annotated and included in the knowledge base. Besides, the issue of qualifying the epistemic nature of the extracted knowledge concerns various domains, not only the biomedical one. It can be seen as related to the trustworthiness and the confidence given to any information found on the Web by humans. Within the Semantic Web community, content-based trust mechanisms still need to be addressed [4]. The categories presented in this paper are mainly discursive categories resulting from scientific discourse analysis in the field of biology. They rely on the more general linguistic (grammatical) categories of enunciation, temporality and modality described and modelled in an ontology [3], meaning that the discursive categories refer to markers from these linguistic categories and thus demonstrate their close interaction.

In this context we[4] worked on the concepts of certainty and uncertainty. Our analysis was performed on three corpora containing respectively ten, twenty and thirteen scientific papers about miRNAs, coming from the specialized database PUBMED. Each of them allowed us to study the biomedical domain, the linguistic phenomenon and to perform manual annotations. From the linguistic analysis, the concept of consolidation emerged and led us to define the articulation of different semantic categories. We then performed a quantitative analysis and experimented an automatic annotation using the GATE platform. We here present the linguistic analysis and the results of the limited but validating evaluation of the semantic categories. Most of the papers in our corpora describe the experiments performed by the authors, others simply present a state of the art in the domain or about a particular

[3] Biologists generally search for information in databases such as PUBMED and MEDLINE

[4] Most of this study was carried out as a Master's Project (DEFI Université Paris X) by A. Folino, P. Geretto, L. Kuznik and M. Younes-Michiels.

phenomenon while others concern only methods. Since the analysis of the third group would require a specific study of their role in the scientific and biomedical area, we decided not to include them in our initial analysis.We focused on papers presenting authors' experiments. The objectives, means, results and phases of the research process are explained in detail in these articles, even if differences appear in the way they are formally organized: not all papers follow the same regular structure of introduction, results, discussion (or results and discussion), materials and methods, and sometimes conclusion.

In the following section, we report on the linguistic phenomenon of epistemic modality that led us to define six semantic categories that are expressed in the experimental biomedical articles of our corpus. In section 3, we present a semantic map organizing these categories. Finally, some perspectives for future research are presented.

2 State of the Art and Methodology

2.1 Background

There is an abundant literature on the use of the linguistic phenomenon of modality, particularly epistemic modality in scientific texts.
Our overview of this phenomenon is based on [5] which analyses the concepts of hedging and modality in relevant previous research such as Hyland's and Markkanen's.

The theoretical basis for this study of epistemic modality and its use in academic discourse is as described in [5]: epistemic modality markers are *"linguistic items that explicitly qualify the truth value of a proposition"*. As for the classification of markers, we will specify the differences and similarities between existing approaches and ours in this section. Concerning the importance of interpreting modality in biomedical texts, [6] affirm that: *"detecting uncertain and negative assertions is essential in most Text Mining tasks, where in general the aim is to derive factual knowledge for textual data. [...] these language forms [...] are intended to express impressions, hypothesised explanations of experimental results or negative findings"*.

Most studies aim at the automatic recognition and extraction of modal expressions and their classification according to particular classes. In [7] the authors propose to classify such expressions according to the type of information they convey: level of certainty, indicating the degree of certainty expressed by the author and including classes such as absolute, high, medium and low; point of view, which distinguishes between the author's and others' ideas; knowledge type, which distinguishes speculations and statements based on experimental evidence, by means of the following classes: speculative, deductive, demonstrative and sensory. The demonstrative class contains verbs such as *demonstrate*, *find*, *show*, *confirm*, etc. As for epistemic modality, [8] identify an axis whose extremes are truly factual and counterfactual events. Between the two extremes there are different modal types: degrees of possibility, belief, evidentiality, expectation, attempting and command. In

[9] the term unhedgers is introduced for verbs such as *demonstrate*, *show*, *prove*, etc. which are considered as conveying a strong degree of certainty in positive sentences and of hedging in negative ones.

2.2 Linguistic Analysis

The analysis of uncertainty in our corpus has required handling the issue of the degree of truth value in certainty. We consider that lexical terms such as *demonstrate, validate, support* [5]… convey a different meaning from *show, find, result…*, hence it was necessary to separate them into two distinct semantic categories. The first one conveys a notion of data strengthening, and the second one conveys a notion of data observation. Since this distinction is one that has not been made by previous linguistic studies, we have to study the concepts of certainty and uncertainty from a new point of view. Our categories are linguistically marked[6] (even if some ambiguities still remain) and semantically linked to different states of scientific knowledge. In order to reduce the lack of knowledge and increase general knowledge, members of the scientific community perform series of experiments from which they draw results that can sometimes confirm one another. Therefore, the notions we propose are: observations, *"consolidation"*[7], objects of study, hypothesis, lack of knowledge and general knowledge. We have also analyzed reported speech, which introduces other authors' knowledge into the articles from referenced sources.

Observations. We have detected within our corpus: *reveal*, *show*, *find*, *report*, *result*, *observe*, *determine*, that are considered as suggesting a high degree of certainty in the above-mentioned studies. Indeed, it is argued in [5] and in several other studies that statements without certainty markers (such as *certainly*…) are more assertive than the same ones containing them. When a writer wants to place the statement beyond doubt he puts a marker. An epistemic marker tends therefore to question the truth-value of the expressed content. But the markers listed above do not convey a weight either of doubt or of certainty. The sentences in which they appear refer to observed phenomena from experiments. The fact expressed with or without any term of the list belongs to direct observation: hencec it is equally as certain, whether introduced by markers or not. Consequently, it is not possible to decide if the author wishes to nuance the truth value of the statement. In our corpus, these markers appear to be simply a stylistic device used to vary the way of expressing facts. Detecting the sentences marked with this kind of specific marker gives a partial view of what is truth-valued. However, the detection of

[5] For ease of reading, the simple forms of the lexical markers are listed.

[6] Table 1 (cf. Appendix) shows the markers found in our corpus.

[7] Consolidation is the translation of the French neologism "Confortation": we think that there is no strictly equivalent term in English, and to our knowledge this semantic notion is introduced here for the first time.

sentences containing these markers is difficult because of the absence of any explicit markers. Hence, not all sentences without markers can be categorized as "observations" and this leads to ambiguity Although it is considered useful to detect these statements, a point confirmed by [10], it will be necessary to further find out a possible way of extracting them.

"Confortation". Terms such as *demonstrate*, *prove*, *confirm…* are present in biomedical texts. These markers express a heightened degree of confidence in the biological statements that are made. Given their role in strengthening the claim, we have designated them by the term *"consolidation"*. The markers of observations are related to experiments, whereas these are related to an idea of validation or confirmation. None of the previously mentioned approaches have attempted this distinction between the two categories. [9] gives the same level of certainty to both. [7] attributes a level of demonstration to what we name *"consolidation"* markers but includes *reveal*, *show* or *find* at that level. *"Consolidation"* clearly marks observations ("*confirming our results")* as well as uncertainty-marked sentences (*"is consistent with the hypothesis"). "Consolidation"* can be internal or external, marking the authors' or others' data.

Objects of Studies and Hypothesis. [5] includes *whether* within the list of lexical uncertainty markers. This word is present in our corpus and expresses the undetermined truth value of a fact. Authors use it to present the questions which they or others set out to answer. It is clearly related to the scope of the investigation: "To address whether …, we first tested …". We have detected that the main ways of expressing the objects of the experiments in our corpus are *whether* and *if*. However another syntactic construction conveys the same meaning: "to verify that …, we cloned …". Further investigation is necessary. Nevertheless, this uncertainty differs semantically from that conveyed by the other uncertainty markers: *suggest*, *believe*, *hypothesis…* or modal auxiliaries such as *may*, *might*, *would*…These markers put forward a speculative idea derived from experiments: *"these two findings suggest that miRs should be ..."*. Authors interpret and give a possible explanation for their observation. Sometimes, observation markers such as *findings* in the above example explicitly indicate that the suggestion comes from experimental results. Even in the absence of these markers, these sentences still convey the same meaning. We therefore propose two semantic categories, namely "objects of study", which represents the aim of the research reported by the authors in the article, and "hypothesis", which corresponds to the reasoning derived from the authors observations during their experiments.

Lack of Knowledge. Expressions such as *"it is still unknown...*", *"it remains unclear..."* etc. are considered as statements conveying lack of knowledge in [11]. These authors attribute to the sentences containing such expressions the lowest level of certainty, i.e. complete uncertainty. Our linguistic analyses confirm that such expressions express lack of knowledge. We do not, however, connect it to

uncertainty. The following sentence is a good example: "*however, their biological functions… remain largely undefined and experimentally untested"*.

General Knowledge. There are some sentences expressing facts, but not related to the study performed by the author nor introduced as another author's point of view. We consider that such sentences convey general knowledge, a truth presented as shared by the scientific community: "*It is currently estimated that the expression of… is…"*. Some expressions like *"it is well known*", *"currently"…* are easy to identify. Moreover, we have observed that, even in the absence of such markers, these sentences are characterized by the use of the present tense: *'RNA silencing (RNAi) is a new gene regulatory mechanism"*. Nevertheless, that criterion is not sufficient to determine unambiguously that an expression belongs to the category in question. The following sequence illustrates the ambiguity:

"Vascular endothelial growth factor A (VEGFA or VEGF) is an essential growth and survival factor for endothelial cells. It plays a major role in physiological and pathological angiogenesis through its ability to stimulate growth of new blood vessels from nearby capillaries (Ferrara 2005). Through alternative splicing, the highly conserved VEGF gene can produce various protein isoforms, with the three principal forms consisting of 121, 165, and 189 amino acids. VEGF is translated from two start codons, each of which is regulated by an independent IRES (Huez et al. 2001)."

The first sentence of the above paragraph contains a verb in the present tense and conveys a meaning that can be considered generic, according to the meaning we have given to general knowledge. The second sentence, however, that presents the same structure and expresses a meaning as general as the first one, contains a bibliographic reference. So, it cannot be certainly affirmed that the first sentence belongs to the dimension of general knowledge, because the authors may have reported both sentences from the cited source, including a reference only after the second one.

Reported Speech. Authors frequently introduce knowledge that refers to experiments from identified research articles. In reported speech, there can be markers of hypothesis, observation or "consolidation": *"It has been suggested that … may involve … [30]."; "VEGF translation has been shown to be… (Akiri et al. 1998); "A recent report posited that…*[5]. *That study demonstrated a region common to the 3'UTRs of..."*. This phenomenon has also been categorized as "point of view" in [7] which points out the difficulty in determining in some contexts whether authors are fully committed to the statements. This happens particularly with reported speech sentences containing markers such as *probably*, *may…* In *'the heterogeneity of miR-34a… was probably due to… as recently reported (Landgraf et al. 2007…)'*, the hypothesis marked by *probably,* is in reported speech. But this does not ascertain whether the authors exactly reported the idea contained in the cited article: the hypothesis could be their own. In the presence of an impersonal subject, as in *"It is suggested that..."*, [6] finds it difficult to detect the paternity of the source and considers that further contextual evidence is required. On the contrary, we tend to consider that these constructions are more likely to convey the source's point of view. But we agree that the fidelity to the reported facts cannot be certified. In [12], the authors consider citations as a common practice in scientific discourse to *"indicate a*

network of mutually supportive or contrastive works", and show that hedging cues are frequent in this context where they have a rhetorical function. According to [5], epistemic modality markers in academic discourse are used in an *"interactive" and "persuasive"* way. The reported speech is introduced by various means such as: bibliographic references; expressions like *other studies*, *recent report*; *authors' names* followed by verbs like *report*; direct citations.

2.3 Semantic Notions Map

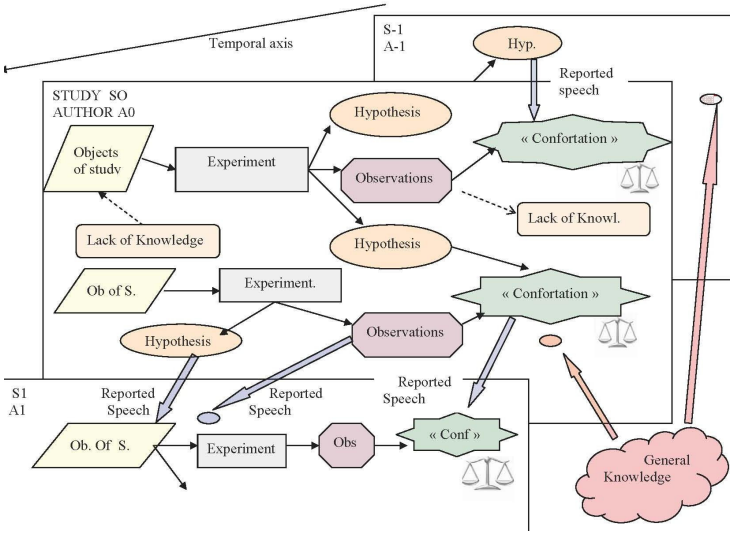

Fig. 1. Study S0 was published at the time T0, S-1 was published previously (time T-1) and S1 afterwards. In the same way, the authors have been identified as A-1, A0 and A1.

The articulation of these notions can be visualized with the example of one possible (inter and intra) discursive situation, as shown in Fig. 1. Notions operate inside an author's study *(S0 in Fig. 1)* and outside it, relating it to other previous (or later) publications *(S-1, S1 in Fig. 1)*. Therefore, they are organized according to a temporal publication axis. Authors express the objects of study they are experimenting on. As explained in the introduction, experiments (the black boxes in the map) are not developed here. Sometimes it is possible to group and compare two or more observations (or hypotheses) from different authors. Authors, in order to support their theories or observations, refer to other hypotheses or to observations from their own experiments, or from others' (*arrows of reported speech in Fig.1*). We can see in Fig. 1 that in study S1, author A1 refers to some observations introduced by "consolidation" markers. Marking with "consolidation" observations that have already been marked with "consolidation" conveys a stronger sense of validation than

marking with “consolidation” observations that are not already marked by “consolidation”. Graphically, this is shown by the different dimensions of the balance. The more observations or hypotheses are compared, the more “consolidation” of observations (or hypotheses) becomes important and acquires weight (*balance in Fig.1*). Authors can also report parts of general knowledge, which is represented as a cloud-form at the bottom right of the map. Each of these semantic categories qualifies the truth value of a proposition differently. Consolidation markers can give a truth value weight to observations as well as to hypotheses, to the authors' as well as to others' data, and to any other consolidation act, thus making it tricky to estimate the qualification of this sort of truth value weight as a validation. Moreover, we think that the final validation of the information and considerations given in each article depends on extra data such as the authors' and laboratories' fame, and the knowledge of the facts described. Final validation rests, in the end, with those who can appreciate the weight of the hypotheses, observations, and consolidations given in the articles.. In the use of a semantic tool processing, we estimate at this stage of our analysis, that the final validation would be given by the curators.

3 Towards automatic semantic annotation of texts

Table 1. Occurrences of the categories in our analysis corpus

Objects of study	*Hypothesis*	*Observations*	*“Confortation”*
to investigate whether	suggest that	Show	Confirm
to identify	Might	Reveal	Support
to test whether	To be likely to	Find	consistent with
to determine whether	potential	be shown	Demonstrate
in order to	should	be reported	provide evidence
to address whether	expected	resulting in	Be confirmed
to assess whether	To hypothesize		In agreement
to explore whether	predict		experimental validation
to evaluate	estimate		Verify
It is still an open question whether	appear		
	assumption		Confirm by demonstrating
to determine if	potentially		similar to
we conducted […] for	possibility		Validate
to ask whether	predictable		In support of
	possibly		
	Would		

Reported Speech + Markers of		
"Confortation"	*Hypothesis*	*Observations*
evidence	May	Show
Support	Estimate that	Be shown to
be supported	be thought to	Reveal that
be demonstrated	studies + suggest	studies+ show
be consistent with	possible	Be found to
/author et al./ + confirm	suggest	indicate that
	might	several reports + indicate that
	would predict	Resulting in
	could potentially	Be revealed + recent studies
	Given[…], would	/author et al/ + show
	be suggest + may	recent report + show
	prediction	Be reported to
		/author et al/ + found

General Knowledge
As assumed currently

Lack of Knowledge
Largely unknown
Remain to be
Be still lacking
To warrant further studies
Unknown
Remain poor
For future research
will be necessary
Unclear
further studies are necessary

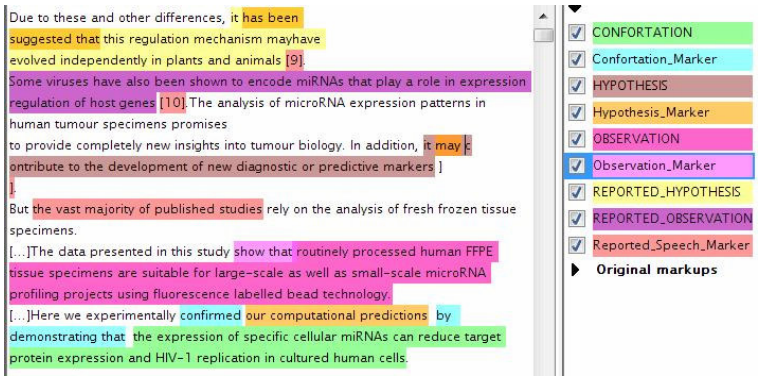

Fig. 2. Manual Annotation within Gate: this picture gives an example of the results we would like to get from the automatic function of Gate

Our automatic annotation tool is still in progress. We have already partially used some categories as keys for extraction patterns in order to automatically annotate texts within the platform GATE. The main limitation of our present approach concerns difficulties in detecting the right scope for each annotation: for sentences with *show*, detection of the proposition is needed; for reported speech, the whole sentence is necessary; in some cases such as for example anaphoric expressions in which *this* or *these* is used, the context before the sentence is essential. We give an example (Fig. 2) of annotations using our semantic categories. We are engaged in the process of testing these lists and enhancing them. However, the difficulty pointed out in section 2 (reported speech) will be to determine whether a hypothesis has been formulated by

the cited author or whether on the contrary, it deals with an idea of the citing author, in which case it would therefore be a reported hypothesis. The table 1 illustrates the classification of markers.

4 Conclusion and Perspectives

In biomedical articles**,** authors give their interpretation of all pieces of knowledge, assigning a truth value weight to each item. In reported speech, we are confronted with the question of determining whom the opinion belongs to. However, the whole article has finally a direction given by its authors. "Consolidation" introduces the notion of reinforcement given to the observations or hypotheses made on biological facts. Deriving from that, the final weight of truth value the authors attribute to each result or speculation is under question. We have observed that "consolidation" can mark the observations and the hypothesis of the authors' own study. This can therefore change the truth value weight assigned to results in the course of their explanations in the paper.

How can the final truth value weight of observations, hypotheses, objects of study, or lack of knowledge be determined? Do the authors alter their initial observations? Do they reduce the lack of knowledge? Do they transform their hypothesis into observed results? Do they answer their objects of study? These points need further investigation. We consider that different states of knowledge can be present in research articles: biological facts can be marked with "lack of knowledge", "object of study", "general knowledge", "hypothesis", "observation" or "consolidation". Among the last three categories, some come from the authors' experiments and some from others' (reported speech). But many points remain to be examined. The main one is the search in sentences without markers for possible linguistic clues that would enable general knowledge to be distinguished from observations. We will also have to explore the methods in biomedical articles. As reported by [11], database curators are often interested in experimental evidence and methods. A monitored evaluation performed with an automatic extraction tool could provide interesting information, giving new directions; furthermore the position in the section of the text (introduction, results, discussion etc.) is a point that will have to be included in further investigation. The use of verbs must also be examined: in particular we have observed that the use of different verb tenses, of negative modal verbs and of the passive voice could have an important role that should be further analyzed.

Lastly, we need to integrate this work with the ontology population process already set up for biologists [2] [3].

References

1. Shotton, D., Portwin, K., Klyne, G. Miles, A., Adventures in semantic publishing : exemplar semantic enhancement of a research articles. (submitted for publication),

2. Jilani, I., Amardeilh, F.: Enrichissement automatique d'une base de connaissance biologique à l'aide des outils du Web sémantique. In: Proceedings of the 20th French Conference on Knowledge Engineering, pp. 169--180. Hammamet, Tunisia, (2009)
3. Battistelli, D., Amardeilh, F.: Knowledge Claims in Scientific Literature, Uncertainty and Semantic Annotation: A Case Study in the Biological Domain. (accepted at KCAP – SAAKM 09)
4. Bizer, C., Oldakowski, R. Using context- and content-based trust policies on the semantic web. In: Proceedings of Alternate Track Papers & Posters of WWW 2004, pp. 228--229 (2004)
5. Vold, E. T.: Modalité épistémique et discours scientifique : une étude contrastive des modalisateurs épistémiques dans des articles de recherche français, norvégiens et anglais, en linguistique et médecine. Doctoral Thesis, University of Bergen, (2008)
6. Svarzas, G., Vincze, V., Farkas, R., Csirik, J.: The BioScope Corpus: annotation for negation, uncertainty, and their scope in biomedical texts. In: Proceedings of BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pp. 38--45. Columbus, Ohio, USA, (2008)
7. Thompson, P., Venturi, G., McNaught, J., Montemagni, S., Ananiadou, S.: Categorizing Modality in Biomedical Texts. In: Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining. Marrakech, Morocco, (2008)
8. Sauri R., Verhagen M., Pustejovsky J.: Annotating and recognizing event modality in text. In: Proceedings of the 19th International FLAIRS Conference, FLAIRS 2006, Melbourne Beach, Florida, (2006)
9. Kilicoglu, H., Bergler, S.: Recognizing speculative language in biomedical research articles: a linguistically motivated perspective, BMC Bioinformatics, 9, (2008)
10. Shatkay, H., Pan, F., Rzhetsky, A.,Wilbur, W.J.: Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. In: Bioinformatics, 24, 2086--2093, (2008)
11. Wilbur, W.J., Rzhetsky, A., Shatkay, H.: New directions in biomedical text annotation: definitions, guidelines, and corpus construction. BMC Bioinformatics 7, (2006)
12. Mercer, R.E., Di Marco, C., Kroon, F.W.: The frequency of hedging cues in citation contexts in scientific writings. In: Tawfik, A.Y, Goodwin, S.D. (eds.). Canadian AI 2004. LNCS, vol. 3060, pp. 75--88. Springer, Berlin Heidelberg (2004)