Knowledge Claims in Scientific Literature, Uncertainty and Semantic Annotation: A Case Study in the Biological Domain

Delphine Battistelli, Florence Amardeilh

Knowledge Claims in Scientific Literature, Uncertainty and Semantic Annotation: A Case Study in the Biological Domain

Delphine Battistelli
MoDyCo – UMR 7114
University Paris X
F92001 Nanterre
+33 140974179
delphine.battistelli@paris-sorbonne.fr

Florence Amardeilh
MoDyCo – UMR 7114
University Paris X
F92001 Nanterre
+33 144923501
florence.amardeilh@mondeca.com

ABSTRACT
Within the framework of a project which includes biologists, computer scientists and linguists, we aim at implementing a knowledge base on a recently discovered biological phenomenon. To ascertain the validity of the knowledge extracted from biological literature and to use it to semantically annotate the contents, we address the temporal, modal and/or enunciative characteristics of the identified information. This work answers a specific need expressed by the biologists regarding the quality of the information extracted and recombined for them in the knowledge base.

Categories and Subject Descriptors
H.3.1 [**Content Analysis and Indexing**]: Linguistic processing.

I.2.4 [**Knowledge Representation Formalisms and Methods**]

I.2.7 [**Natural Language Processing**]: Language model.

General Terms
Design, Experimentation, Standardization, Languages, Theory.

Keywords
Modality, hedging, ontology population, semantic annotation.

1. INTRODUCTION AND CONTEXT
Within the framework of the Microbio project, a collaboration between biologists, computer scientists and linguists, we aim at implementing a knowledge base on a recently discovered biological phenomenon which was awarded the Nobel Prize in 2006. This phenomenon concerns miRNA, their mutation and regulation and their impact on the development or inhibition of certain diseases. The knowledge base collects data from the existing distributed databases but also from the biomedical literature available on the PubMed portal. This knowledge base is intended to be consulted by biologists but also to contribute to the semi-automatic annotation of biomedical literature in this specific domain.

An initial domain ontology about miRNA was modeled by taking as input various existing resources (such as the databases, the Sequence Ontology) and a set of expert interviews with biologists from the Pasteur Institute in Montevideo. As described in [1], an information extraction tool aiming at populating the miRNA ontology was configured based on a corpus of full text articles from Medline selected for their relevance and domain coverage by the biologists. It can automatically identify and tag sentences which state a miRNA and at least a Gene and/or a disease and/or a Mutation (SNP). The CA Manager infrastructure [2] was used to configure and develop the service that enriches the knowledge base with new instances tagged by the information extraction tool. This tool was chosen because it already includes a set of consolidation algorithms that can automatically disambiguate some information, resolve URIs and check domain, range and restriction constraints based on the ontology model of the knowledge base.

However, to ascertain the validity of the new knowledge and to use it to semantically annotate content, the biologists pointed out that this is not enough. In fact, temporal, modal and/or enunciative features indicative of authorial commitment to the identified information in biological texts need to be explicitly tagged and included in the knowledge base. This aspect of our approach constitutes its originality and answers a specific need expressed by the biologists regarding the quality of the information extracted and recombined for them in the knowledge base: namely, the need to have access to the validation framework of the information (*in vitro, in vivo, in silico* … experimentations, by which team, when, the certainty or uncertainty status of the results, and so on). Within the biological domain, [3] suggest that "contradictions and speculations in the literature are likely to prove a fruitful source of new hypotheses. All of this is territory yet to be explored".

We address the issue by focusing on the explicit linguistic markers of epistemic qualification; that is to say by working directly on the textual units that express a certain degree of confidence about a new item of knowledge. These textual units are often classified in the linguistic literature under the term "modality" or "hedging" and have been widely explored in scientific corpora. None of these previous approaches, however, have attempted a formal representation of this linguistic category.

In this article, we present our preliminary work for establishing a conceptual linguistic framework by the means of an ontology of what we name "the linguistic context of validation" of an item of information. This ontology aims to capture the elements of the information perspectivization expressed by the author of a text when using certain modal markers (*may, maybe*, …), references to other enunciators' claims or tenses (e.g. continuous present). Thus, this ontology represents the relations between three

linguistic categories involved in the characterization of validity conditions of information: enunciative conditions, aspecto-temporal conditions and modal (or rhetorical) conditions. These categories are relevant for any domain. Thus the ontology is independent of any domain ontology and reusable for other applications. In this paper we particularly focus on the analysis and the modeling aspects of modal and/or enunciative characteristics of tagged sentences.

This ontology can be used to semantically annotate the extracted information for a subsequent ontology population. In the next section, we discuss how this need for extra information about new knowledge also concerns the idea of trust in knowledge acquisition and semantic web fields. We then present the linguistic perspective, i.e. how modality is expressed in texts, and our model for an ontology of modality. In section 4, we illustrate our approach within the Microbio project on the miRNA phenomenon to explain how it can be used to semantically annotate textual information in order to enrich a knowledge base.

2. TRUST IN KNOWLEDGE ACQUISITION

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.

Current knowledge acquisition tools aim to automatically or semi-automatically produce semantic annotations or new knowledge instances from textual documents, based on an ontology. If we look at the Semantic Web architecture proposed by Sir Tim Berners-Lee [4][5], it is composed of several layers that target specific needs for putting into practice this vision of a "semantic web", interpretable by both humans and machines. Among these layers, the trust layer is one of the top semantic web architecture layers still awaiting standards, or at least recommendations to the community. This layer proposes to answer the following questions: how trustworthy is information found on the semantic Web? How do I decide that it is trustworthy? In 2004, [6][7] defined three types of trust mechanisms:

- Reputation-based Trust Mechanisms
- Context-based Trust Mechanisms
- Content-based Trust Mechanisms

While the first two types refer more to security issues and have been studied since 2003 by the Semantic Web community in different working groups, the last one still needs to be addressed. This is particularly the case when semantically annotating a web resource and populating an ontology from the analysis provided by information extraction tools. How can the content itself be used to characterize the trustworthiness of the extracted information to be added to the knowledge base? This can be further redefined as addressing the following issues:

- aggregation of the same content published by different information providers
- following the same event or knowledge evolution over a certain timeline (e.g. from rumors to announcements to facts in the competitive intelligence domain)
- fine-grained characterization of the semantics of the information itself: the possibility that an event may occur and how sure are we of its occurrence?

The first two issues can be addressed using consolidation algorithms based on inference rule or constraint validation from a reasoning engine plugged to the ontology repository [9]. On the other hand, the last one requires more information to be extracted from the text in order to guide the end-user about the trustworthiness of the acquired knowledge. In fact we try to answer the following question: how can we identify in a textual document "valid" knowledge which can be used to annotate the resource and be stored in a knowledge base but also queried and trusted by the end-user. To address that question, we plan to look at the textual modalities that can be found in the text and that can be linked to the domain-related knowledge instance itself.

3. EXPLORING MODALITY OR HEDGING IN SCIENTIFIC TEXTS

3.1 Epistemic modality in texts: state of the art and presentation of our methodology

The named entity and relation identification tasks are common in the Text Mining field applied to biological textual corpora [9]. The task that consists in identifying the modality occurrences and giving them, if possible, a semantic interpretation in their context of occurrence is less classical and represents a challenge for the Text Mining field in general, and no doubt in the biomedical domain as well. For [10], "detecting uncertain and negative assertions is essential in most Text Mining tasks where in general, the aim is to derive factual knowledge from textual data. This is especially so for many tasks in the biomedical (medical and biological) domain, where these language forms are used extensively in textual documents and are intended to express impressions, hypothesised explanations of experimental results or negative findings." The first research attempt that aimed at exploring the area of modality in biomedical texts was [11]. It is now possible to distinguish major approaches depending on whether or not they refer to extensive linguistic analysis of modality in scientific texts as in [12][1]. This is the case for [13] and [17] but not for [11], [14] or [15]. The latter approach consists in defining manually a set of annotation schemas to be used mainly in machine learning algorithms[2].

Our approach is closer to the methodological and conceptual one of [13], [16] or [17] since we explicitly integrate in our modeling work the fact that the semantic – and/or pragmatic – interpretation of the modal markers still remains an open issue, unresolved in a

[1] It should be mentioned that most of the analyses have been done on English. Two exceptions, however, are [18], which provides a comparative epistemic modality analysis of scientific corpora in three different languages (French, English and Norwegian), and [19], which presents the problem of epistemic vs. rhetorical interpretation of modal markers in French scientific discourse.

[2] See [9] for a presentation of the limits of this type of approach (called 'corpus linguistics') in terms of linguistic semantic analysis.

systematic way. This is essentially due to the following three factors:

(i) no consensus has yet emerged regarding the types and numbers of semantic values to be retained (on a scale from 'totally certain' to 'totally uncertain') [17];

(ii) the problem of the ambiguity of these markers in the textual context is raised (i.e. their precise value or their domain of semantic coverage) [18];

(iii) lastly, the main ambiguity resides in their semantic *vs.* pragmatic interpretation (these markers work as many clues for epistemic or for rhetorical functions) [12] [19].

As we will see below, the fact that these factors are integrated from the outset in our modeling proposal allows us to consider an operative automatic annotation tool for modality just as it is expressed in the texts.

Several other points deserve to be stressed concerning our methodology and our modeling process for annotating modality in texts.

First of all, in the exploration and analysis of modal markers in a scientific textual corpus, our strategy involves selecting a set of sentences that are already relevant for the biologists (e.g. a set of sentences about relations between miRNA, genes and diseases for instance). It should be remembered that one of the methodological principles of our modeling task consists in starting from a need and its analysis (the one formulated by the Pasteur Institute biologists) and thus from a domain and a specific application objective. This methodological position is based on a prerequisite formulated by the Knowledge Engineering community when creating ontologies [20]. We thus propose a different view from studies interested in the expression of modality in texts, as these studies generally analyse all the sentences of a text (within the biomedical domain or not, cf. works cited above). Moreover, we contend that our approach simulates more accurately the reading process of scientific researchers in their disciplinary field], or, to put it differently, replicates the process whereby entities (and/or the relations between entities) are first identified and then the linguistic characteristics of validation modes are considered (information presented as certain or not, by who, when…).

Second, another divergence with the above cited works is that we envisage the definition of a linguistic ontology of modality, trying to take advantage of the inference rules that an ontology permits with a view to placing the information retrieval task of a biologist towards the possible formulation of new hypotheses or discoveries by displaying the textual fragments organised according to epistemic criteria. It should be mentioned here that although projects for creating linguistic ontologies already exist (see [21][22] and [23] on the biomedical domain in particular), none address the issue of creating an ontology of modality. Furthermore, these projects underline the specific difficulty in creating an ontology for categories that are not lexical but grammatical (such as time, aspect, modality, gender…). They handle the issue of creating linguistic ontologies referring not to a grammatical category but to the rhetorical structure of texts ([20][21]). However, [23], which is closer to our approach, even if the authors do not mention that they are interested in the grammatical category of modality, questions the distribution of textual units such as 'research statements', 'research questions' and 'comments' in the texts.

Third, and this point distinguishes our work from studies conducted more generally in the field of modality (with the exception of the annotation standard TimeML [24]), we propose to link modality explicitly to the categories of enunciation and temporality. This choice appears in the definition of our ontology as the root concept is 'Linguistic_Context_of_Validation". Not only does this approach have a sounder theoretical basis in linguistic analysis (cf. general linguistics studies that mention the close interweaving between these categories), it also has a clear applicative interest in terms of activity for competitive intelligence tasks in science. It consists in being able to manage the evolution overtime in the validation status of the information units that are being monitored as well as their characteristics in terms of enunciative responsibility (by whom – either an individual or a collective scientific entity – a scientific fact is claimed in a text). That is why we speak about the "context of linguistic validation" of an occurrence of a fact reported in a text. This concept has three subtypes: "modal context", "enunciative context" and "temporal context". In the remainder of this paper, we focus mainly on the "modal context" and the "enunciative context" by analysing a relevant sentence from our biomedical corpus.

3.2 The ontology of Modality

The concept "Linguistic_Context_of_Validation" fits into a larger specification of a context of validation of an occurrence of an identified fact. Consequently, the root concept of our ontology is "Context_of_Validation" which has two sub-concepts, the Linguistic one and an Extra-linguistic one.

The "Linguistic_Context_of_Validation" deals with the knowledge that is relative to the context of validation as expressed by the author of the text, whereas the "Extra_Linguistic_Context_of_Validation" looks at external knowledge of a text (such as its Impact Factor). The latter concept will not be discussed here.

Taking into account the enunciative aspect enables us to provide one part of the linguistic context of validation of the propositional content (more prosaically called 'identified fact' here)[3], presented with the following sub-concepts as shown later in Figure 2:

- "Full authorial commitment", i.e. the author of the text accepts full responsibility for the claims made in the text[4]:

Conference'04, Month 1–2, 2004, City, State, Country.

[3] The essential parameters of the state of enunciation are "by whom", "when", "where" and "to whom it was said". In this paper, we address only the parameters "by whom" and "when".

[4] By default, the author of a text takes responsibility for the entire set of statements of a text, that is to say that he is the primary enunciative source. This conception is explicitly adopted in

- "Indirect Speech", i.e. attributed to another enunciator, with a marker of agreement or on the contrary of distance from the author;
- "Limited temporal universe", e.g. "X said in 2002 that…";
- "Limited experimental universe", e.g. "in the experimentation…, it appears that….";
- "Limited thematic universe", e.g. "Regarding…, X said that…";

Taking into account the modal (epistemic) aspect[5] enables us to provide another part of the linguistic context of validation of the propositional content, represented by the following sub-concepts:

- "No_Doubt", e.g. "X regulates Y";
- "Doubt", e.g. "X may regulate Y";

These sub-concepts can be extended with a more fine-grained scale of the degree of modality expressed in a text.

Taking into account the temporal aspect[6] enables us to provide yet another part of the context of validation of the propositional content, represented by the following sub-concepts:

- "Has happened", i.e. having occurred at a date or over a period of time (not necessarily explicitly mentioned in the text);
- "In progress", i.e. being developed at a date or over a period of time (not necessarily explicitly mentioned in the text);
- "Stable", e.g. established as a stable truth at a date or over a period of time (not necessarily explicitly mentioned in the text).

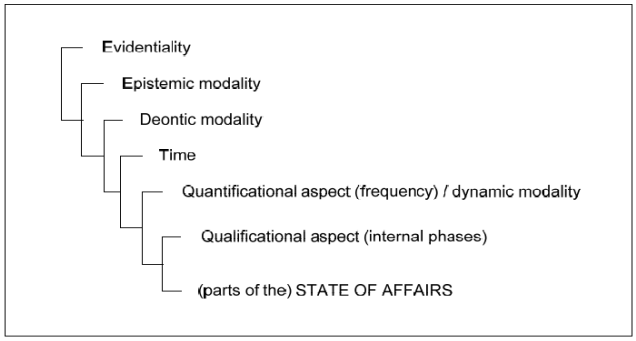

Figure 1. Hierarchy of constitutive operations of a statement according to Nuyts [27]

work within the French enunciative linguistics field (Bally, Benveniste or Culioli) and also in Anglo-Saxon computational linguistics, for instance in [24] and [25] looking at the embedded discourses or 'nested-sources' analysis.

[5] In fact, following [26] for instance, we only handle the epistemic aspect of modality (the one relative to the degree of certainty expressed regarding the truth of a propositional content) which directly concerns our approach to analysing scientific texts.

[6] As the categories of time and aspect are interrelated from the point of view of linguistic analysis, we have preferred to explicitly label the concept as such.

These three key aspects refer to part of the hierarchy of the constituent operations of a statement presented in [27] – cf. Figure 1. Many linguistic analyses postulate the existence of an ordered organization of the constituent operations of a statement, in which the enunciative aspect occurs more or less explicitly. While there is some debate about certain specific features of this organization, it reflects nevertheless a widely-accepted trend towards a hierarchy of temporal, aspect and modal-based operations in the construction of a statement. We note that the markers for these categories do not have the same impact and assume the hierarchical organization described in **Figure 1**. The category of "Evidentiality" that marks the source of information ("according to what I know", "as I see it"…) has the largest range.

In the current state of our work, the linguistic ontology of the linguistic context of validation of an occurrence of a fact is presented in Figure 2[7].

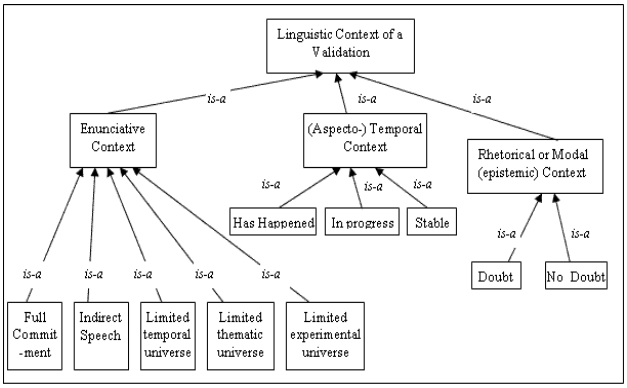

Figure 2. Hierarchical concept tree of the ontology of modality that permits to provide the linguistic context of validation of an occurrence of a fact in scientific texts

Regarding point (i) (cf. section 3.1. above), experiments conducted to measure inter-annotator agreement on the basis of a fine-grained scale of certainty reveal the difficulty of specifying the analytical categories. Recent work by Rubin [17] concludes that "*it is not entirely clear that a five-level distinction of certainty [ABSOLUTE, HIGH, MODERATE, LOW CERTAINTY and UNCERTAINTY] is preferable to a simplistic distinction between statements with certainty and statements with doubt*".

Regarding point (iii) (cf. section 3.1. above), it is worth recalling the difficulties underlined by several studies, among them those mentioned by Rubin [17]: "*Little attempt, however, has been made in natural language computing literature to manually annotate and consequently automate identification of statements with an explicitly expressed certainty or doubt, or shades of epistemic qualifications in between. This lack is possibly due to the complexity of computing epistemic interpretations in different pragmatic contexts; and due to unreliability of variety of*

[7] As in [26], we chose to separate evidential and modal categories. This is an important theoretical option to point out and it distinguishes our approach from that of [25] for instance, where, as in [27], the category of 'evidentiality' is considered as a sub-category of 'modality'.

linguistic expressions in English that could explicitly qualify a statement. Another complication is a lack of agreed-upon and easily identifiable discrete categories on the continuum from certainty to doubt".

Corroborating the remarks made by [13], "*Hedging is critical in scientific discourse because it helps gain communal acceptance for knowledge. Scientific 'truth' is as much a social as an intellectual category, and the distinction writers make between their subject matter and how they want readers to understand their relationship to is crucial to such a highly self-conscious form of discourse. Not only does it influence the effectiveness and credibility of argumentation, but helps define what it means to write science…*".

4. THE MICROBIO PROJECT: AN ILLUSTRATION FOR THE USE OF THE ONTOLOGY OF MODALITY

4.1 Microbio ontology

The Microbio ontology aims at formalizing the recent discoveries related to miRNAs: their regulation with the genes, and especially the mRNA, and their mutations which can cause or on the contrary inhibit diseases. Although no ontologies specifically represent this domain, several terminological and ontological resources, such as the Gene ontology [29] or the Sequence Ontology [30], propose a general overview of genes and their sequences. Some databases on mRNAs and miRNAs, such as Tarbase [31] and miRBase [32], have recently been created, testifying to the interest of biologists in this new issue, which was awarded the Nobel Prize in 2006.

For this project we needed a high-level ontology that would be the starting point for modelling a new ontology dedicated to the representation of miRNAs and their impact on gene regulation and mutation. We therefore decided to work with the Sequence Ontology (SO) as it already represents biological sequences in general and especially the concepts of "Gene", "mRNA" and "miRNA". Some interesting relations between concepts are also described such as « is_part_of » to represent the decomposition of a miRNA into its atomic parts (« loop », « stem »…) or "regulated_by" to model the phenomenon of regulation between a miRNA and a mRNA segment even if it has the status "deprecated" in the Sequence Ontology. Finally, we enriched and extended this subpart of the SO with current knowledge about miRNA thanks to a series of interviews with domain experts, namely the biologists at the Pasteur Institute in Montevideo.

We modelled the Microbio ontology in the OWL as it allows us to keep a reference to the SO concepts reused from the SO and therefore to ease the integration of our ontology in the SO if relevant. When a concept comes from the SO or is equivalent to an existing concept in the SO, then a semantic link between the Microbio concept and the SO one is created via the "owl:equivalentClass" construction. For instance, the concept "miRNA" from the SO is an equivalent class to the concept "miARN" of the Microbio ontology (in French so far). We apply the exact same principle for equivalences between properties thanks to the "owl:equivalentProperty" construction. By initially adopting a separate representation from the SO, the conceptual independence of our ontology can be maintained until it has been definitively validated by the biologists, while still keeping a link to the SO for the time when we will ask the SO working group to integrate our work in their ontology.

The goal of the Microbio ontology is to be automatically instantiated from a corpus of scientific articles extracted from the Medline portal and dealing with miRNAs. A semantic annotation platform named CA Manager [2][33], based on both UIMA and Semantic Web standards, has been used to set up an ontology population workflow, including an information extraction tool dedicated to the identification and tagging of the miRNA, the mRNA, the genes, the disease and the mutations (SNP code) in the text [1] with a Sesame server to store the discovered knowledge instances of both concepts and properties. Another benefit in using the OWL format is that some basic inference rules can be applied to the knowledge base in order to detect new relations between miRNAs and genes or even between miRNAs and diseases. For the biologists, however, even detecting new relations is not enough if they do not know the degree of certainty and therefore of confidence that they can attribute to the information so as to process it in their own biological analysis.

4.2 Applying modalities to Microbio ontology

To overcome this issue, we proposed the ontology of modality to the biologists and we integrated it with the Microbio ontology. The ontology of Modality is also formalized in OWL and we modeled the following object property "modal:validates(modal:Linguistic_Context_of_Validation, owl:Class)" that allows us to describe any other class with an instance from the concept 'Context of Validation' or one of its sub-concepts. This kind of representation is flexible enough for the ontology of Modality to be used to annotate any other domain ontology. Other concepts and properties have been added to the ontology of Modality to represent the textual context, such as:

- modal:marks(modal:Textual_Unit, modal:Linguistic_Context_of_Validation),
- modal:has_origin(modal:Indirect_Speech, modal:Source),…

To enrich the semantic graph produced by the CA Manager platform in order to automatically populate the Microbio ontology with the modality context, we adopted the following method:

i. First, in a corpus composed of scientific full texts obtained from PubMed with the query [miRNAS+human], we isolate the sentences that carry information about an explicit relation between instances of two different concepts: one (or several) miRNAs and one (or several) genes. This is done through the use of patterns such as $miRNAS+$Gènes. A later step will consist in adding a list of relational markers (verbal, nominal or adjectival) between these two entities, to be defined by the Pasteur Institute biologists.

ii. Second, we work on this set of sentences and analyze their modal markers.

Let's take an example and look at the following sentence:

"The earlier work of Anttila et al. (2003) has also suggested the role of DNA methylation in CYP1A1 regulation".

The blue part of the sentence refers to the Microbio ontology whereas the red part refers to the ontology of Modality, expressing indirect speech, e.g. Anttila et al. (2003), and the uncertainty of this regulation by the use of the verb "suggest". This sentence can be represented by the semantic network as illustrated in Figure 3.

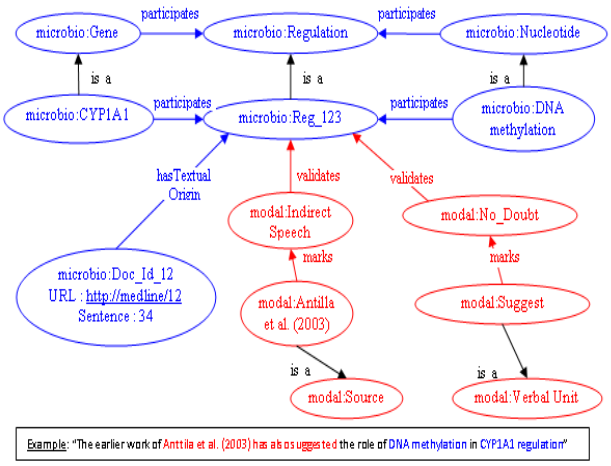

Figure 3. An example of a semantic network including modality information

This semantic network is consolidated and disambiguated by the algorithms provided by the CA Manager. These algorithms automatically control the ontology constraints (domain, range, restrictions…) and resolve the reference to existing instances in order to avoid any redundancy within the knowledge base. A link to the original document is kept for each instance created in the knowledge base so that the scientific articles in which the information was found are easily retrieved for later use and validation. The whole OWL graph is then imported and stored in the repository, ready to be manually validated by the biologists if necessary and queried by both the biologists and the inference engine.

5. CONCLUSION AND PERSPECTIVES

According to [34], "key aspects are to discover unsuspected, new knowledge hidden in the vast scientific literature, to support data driven hypothesis discovery and to derive meaning from the rich language of specialists as expressed in the plethora of textual reports, articles, etc. With the overwhelming amount of information (~80%) in textual unstructured form and the growing number of publications, an estimate of about 2.5 million articles published per year [35] it is not surprising that valuable new sources of research data typically remain underexploited and nuggets of insight or new knowledge are often never discovered in the sea of literature. Scientists are unable to keep abreast of developments in their fields and to make connections between seemingly unrelated facts to generate new ideas and hypotheses". Following the experiment conducted by [36] and described in [13], we aim at providing a tool to support biologists in their daily tasks for annotating scientific articles and (semi-)automatically populating a knowledge base dedicated to a growing field of research. This tool collects textual units scattered in the scientific literature about a particular theme (here the relations between miRNAs and genes) according to an epistemic criterion linked to the presence of certain modal markers in the relevant sentences.

The next step for the project is to evaluate this ontology of Modality and the benefits that the biologists could actually get from. Then we will configure the semantic annotation tool with the patterns that will allow automatic identification of the enunciative, modal and aspecto-temporal concepts in the scientific articles. Then, the whole ontology population application can be implemented and evaluated by the biologists from the Pasteur Institute in Montevideo.

We would also like to participate in the construction of the Trust layer in the Semantic Web architecture and combine our approach with the use of named graph for the enrichment of the semantic network generated on the domain from the textual resources by the modality knowledge. We could then propose an annotation scheme to identify and implement a set of Content-based Trust Mechanisms.

6. ACKNOWLEDGEMENTS

The Microbio project is partially funded by the Stic-Amsud collaborative research program.

7. REFERENCES

[1] Jilani, I., Amardeilh, F. 2009. Enrichissement automatique d'une base de connaissances biologiques à l'aide des outils du Web sémantique. In proceedings of the 20th French Conference on Knowledge Engineering, Hammamet, Tunisia

[2] Amardeilh, F. 2008. Semantic Annotation and Ontology Population. Semantic Web Engineering in the Knowledge Society, ISI Global.

[3] Wilbur, W.J., Rzhetsky, A., Shatkay. H. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinformatics 7 (356).

[4] Berners-Lee, T., Hendler, J. and Lassila, O. 2001. The Scientific American Magazine. DOI= http://www.sciam.com/article.Cfm?Id=The=Semantic-Web

[5] Berners-Lee, T. 2006. Artificial Intelligence and The Semantic Web. DOI= http://www.w3.org/2006/Talks/0718-aaai-tbl/Overview.html#(14)

[6] Carroll, J.J., Bizer, C., Hayes, P., Stickler P. 2005. Named graphs, provenance and trust. In Proceedings of WWW 2005, 613-622

[7] Bizer, C., Oldakowski, R. 2004. Using context- and content-based trust policies on the semantic web. In Proceedings of Alternate Track Papers & Posters of WWW 2004, 228-229

[8] Amardeilh, F., Carloni, O., Noel, L. 2006. PressIndex: a Semantic Web Press Clipping Application. In Proceedings of the ISWC 2006 Semantic Web Challenge, Athens, Georgia, USA, 5-9 November 2006.

[9] Ananiadou, S., McNaught, J. (eds). 2006: Text Mining for Biology and Biomedicine. Boston and London: Artech House.

[10] Szarvas, G., Vincze, V., Farkas, R., Csirik, J. 2008. 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, 38–45, Columbus, Ohio, USA, June 2008.

[11] Light, M., Qiu, X.Y., Srinivasan, P. 2004. The language of bioscience: facts, speculations, and statements in between. In Proceedings of BioLINK 2004: Linking Biological Literature, Ontologies and Databases 2004, 17-24.

[12] Hyland, K. 1998. Hedging in scientific research articles. Amsterdam, Netherlands: John Benjamins B.V.

[13] Mercer, R.E., Di Marco, C, Kroon. 2004. The frequency of hedging cues in citation contexts in scientific writing. LNCS 3060.

[14] Medlock, B, Briscoe, T. 2007. Weakly supervised learning for hedge classification in scientific literature. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics 2007, 992-999.

[15] Thompson, P, Venturi, G, McNaught, J, Montemagni, S, Ananiadou, S. 2008. Categorising modality in biomedical texts. In Proceedings of the LREC 2008, Workshop on Building and Evaluating Resources for Biomedical Text Mining 2008.

[16] Mc Enery, T., Wilson, A. (eds.). 1996. Corpus Linguistics, Edinburgh: Edinburgh University Press.

[17] Rubin, V.T. 2007. Stating with Certainty or Stating with Doubt: Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements. In Proceedings of NAACL HLT 2007, Companion Volume, 141–144, Rochester, NY, April 2007

[18] Vold, E.T. 2008. 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. Université de Bergen.

[19] Latour B. 2001 Le métier de chercheur. Regard d'un anthropologue. 2eme édition revue et corrigée. INRA Editions.

[20] Bachimont, B. 2000. Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances, In J. Charlet, M. Zacklad, G. Kassel & D. Bourigault (Eds.), Ingénierie des connaissances, évolutions récentes et nouveaux défis. Paris, Eyrolles, Chap. 19, 305-324.

[21] Goecke, D., Lüngen, H., Sasaki, F., Witt, A. Farrar, S. 2005. GOLD and Discourse: Domain- and Community-Specific Extensions. In: Proceedings of the E-MELD Workshop on Morphosyntactic Annotation and Terminology: Linguistic Ontologies and Data Categories for Language Resources. Cambridge, Massachusetts.

[22] Bärenfänger, M., Hilbert, M. , Lobin, H., Lüngen, H., 2006. Using OWL ontologies in discourse parsing, In: Kühnberger, Kai-Uwe / Mönnich, Uwe (eds.), Proceedings of the Workshop of Ontologies in Text Technology. Osnabrück.

[23] Ciccarese, P., Wu, E., Clark, T. 2007. An Overview of the SWAN 1.0 Ontology of Scientific Discourse. In Proceedings of the 16th International World Wide Web Conference (WWW2007). Banff, Canada. May 8-12.

[24] Pustejovsky, J., Ingria, R. , Saurı, R., Castaño, J., Littman, J., Gaizauskas, R., Setzer, A., Katz, G., Mani, I. 2005. The specification Language TimeML, in Mani, I., Pustejovsky, J., Gaisauskas, R. (eds) The Language of Time. A Reader, Oxford Linguistics, Oxford University Press Inc., New York: 545-558.

[25] Wiebe, J., Breck, E., Buckley, C., Cardie, C., Davis, P., Fraser, B., Litman, D., Pierce, D., Riloff, E., Wilson, T., Day, D., Maybury, M. 2003. Recognizing and Organizing Opinions Expressed in the World Press. In proceedings of AAAI Spring Symposium on New Directions in Question Answering, 2003

[26] Sauri, R., Pustejovsky, J. 2007. Determining Modality and Factuality for Text Entailment. IIn Proceedings of ICSC 2007, September 17-19, 2007, Irvine, California, 509-516.

[27] Nuyts, J. 2006. Modality : overview and linguistic issues, in Frawley, W. (ed) The expression of modality, Berlin, Mouton de Gruyter, 1-26.

[28] Palmer, F. R. 1986. Mood and Modality, Cambridge, Cambridge University Press, Ed. 2001.

[29] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 25(1), 25-29.

[30] Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R. 2005. The Sequence Ontology: A tool for the unification of genome annotations. Genome Biology, 6(5).

[31] Sethupathy, P., Corda, B., Hatzigeorgiou, A. G. 2006. TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA, 12, 192-197

[32] Griffiths-Jones, S. 2004. The microRNA Registry. Nucleic Acids Research, 32(Database Issue), D109-D111.

[33] Martin J., Herrero G., Capellini A, Francart T., Amardeilh F., Marinova Z. 2008. TAO Suite: Architecture and integration requirements and specifications, TAO project, Deliverable D5.2.

[34] Ananiadou, S. 2009. Text Mining for Biomedicine. In Violaine Prince and Mathieu Roche (eds), Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration, IGI Global. 1-9.

[35] Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Stamerjohanns, H. and Hilf, E. 2004. DOI= The Access/Impact Problem and the Green and Gold Roads to Open Access. Serials review, 30 (4).

[36] Blaschke, C., Andrade, M.A., Ouzounis, C., and Valencia, A. 1999. Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions. In Proceedings of International Conference on Intelligent Systems for Molecular Biology (ISMB 1999), 60–67.

[37] Mercer, M., Marco M. 2004. A Design Methodology for a Biomedical Literature Indexing Tool Using the Rhetoric of Science. In Proceedings of HLT-NAACL 2004 Workshop Biolink 2004, Linking Biological Literature, Ontologies and Databases, 77-84.