Identifying Claimed Knowledge Updates in Biomedical Research Articles

Ágnes Sándor
Xerox Research Centre Europe
Agnes.Sandor@xrce.xerox.com

Anita de Waard
Elsevier Labs, USA
A.Dewaard@elsevier.com

Abstract

Key knowledge components of biological research papers are conveyed by structurally and rhetorically salient sentences that summarize the main findings of a particular experiment. In this article we define such sentences as Claimed Knowledge Updates (CKUs), and propose using them in text mining tasks. We provide evidence that CKUs convey the most important new factual information, and thus demonstrate that rhetorical salience is a systematic discourse structure indicator in biology articles along with structural salience. We assume that CKUs can be detected automatically with state-of-the-art text analysis tools, and suggest some applications for presenting CKUs in knowledge bases and scientific browsing interfaces.

1 Introduction

Biomedical research articles describe newly discovered biological findings, and in doing so, update the readers' knowledge on a particular topic. These two functions of research articles – describing reality and updating knowledge in a field – mobilize different forms of linguistic expression: on the one hand, in order to describe pieces of reality, the authors refer to biological objects and relationships among them, and on the other hand, they shape the way in which new knowledge is inserted into existing accumulated knowledge, through argumentation, discourse and rhetorical structure. The designers of text mining systems are increasingly aware of the importance of integrating both aspects into annotation schemes, and thus models of argumentation, discourse and rhetorical structure are becoming integrated with models of biological reality in modern annotation systems, such as described in Liakata et al. (2010), Nawaz et al. (2010), Wilbur et al. (2006), Sándor (2007), Teufel (1999) and Collier (2006).

Models of biological knowledge are commonly mapped to well-defined linguistic elements like named entities (mostly noun phrases), relationships between the entities (mostly predicates), and these are reliably detected with state-of-the-art text-mining tools (e.g., Nawaz et al. 2010). But the detection of argumentation, discourse and rhetorical structures, and the association of linguistic expressions with these elements, is far less straightforward. The great number of proposed approaches already makes it clear that it is difficult to provide easily applicable and generally accepted annotation guidelines, which can easily be implemented in a web-based environment. An ideal discourse annotation system would be straightforward to use, and it would not require any learning – in the same way that using hyperlinks is a straightforward way to create references. Such an annotation model should also provide a substantial improvement to users who want to find relevant new knowledge.

Here, we propose a simple discourse annotation model to detect the main new knowledge claims in biology research papers. We also propose some suggestions for the implementation of the automatic detection of this model.

2 Claimed Knowledge Updates

Biomedical articles contain a great number of biological propositions, but not all of them are equally relevant: some are central claims, while others merely support the findings; some are factual, while others are merely hypothesized. The authors often summarize their main findings in the

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 10–17,
Jeju, Republic of Korea, 12 July 2012.

title, section titles and caption titles. In addition to these – structurally defined – summaries, the authors also formulate their main findings in rhetorically salient sentences. This rhetorical salience is conveyed via metadiscourse, by which the authors explicitly attribute the findings to themselves, and state that they are based on the current empirical work, such as: "*Our results demonstrate*", "*In the present study we identified*". We will call biological propositions summarized in such structurally or rhetorically salient sentences Claimed Knowledge Updates (CKU).

We hypothesize that a listing of the CKUs in a paper constitutes new main knowledge provided in that paper, and thus we propose that their detection may play an important role in text mining.

We define CKUs as follows:

1. A CKU expresses a verbal or nominal proposition about biological entities.
2. A CKU is a new proposition.
3. The authors present the CKU as factual.
4. A CKU is derived from the experimental work described in the article.
5. The ownership of the proposition is attributed to the author(s) of the article.
6. 4) and 5) are either explicitly expressed or are implicitly conveyed by a structural position as title, section or caption title.

Sentence	CKU
Here we used mass spectrometry to **identify**HuD as a novel neuronal SMN-interacting partner.	HuD is a neuronal SMN-interacting partner.
Our analysis of known HuD-associated mRNAs in neurons **identified** cpg15 mRNA as a highly abundant mRNA in HuDIPs compared with other known targets of HuD, such as GAP43 and Tau.	cpg15 mRNA is a highly abundant mRNA in HuDIPs
Our finding that SMN protein associates with HuD protein and the HuD target cpg15 mRNA in neurons led us to ask whether SMN deficiency affects the abundance or cellular distribution of cpg15 mRNA.	SMN protein associates with HuD protein
	SMN protein associates with cpg15 mRNA

Table 1. Sentences and CKUs from Akten et al.

As an example, Table 1 contains some CKUs from an article on Spinal Muscular Atrophy (Akten et al., 2011). The metadiscourse indicating CKUs is given in bold.

In studying this paper, we found a striking regularity in the appearance of CKUs throughout the article: the Abstract, the Introduction, the Results and the Discussion sections are repeat the same CKUs, as follows:

- in the Abstract they appear as a list of findings;
- in the Introduction, they are inserted within the context of previous knowledge;
- in the Results section, they are explained within the context of the authors' work, and thus provide empirical evidence;

and finally,

- in the Discussion, they are presented in the perspective of the advances in the research domain.

In other words, the four predefined structural units of research articles give an indicator of the underlying CKU organization. This regularity shows that rhetorical salience is systematically related to structural organization, and thus that the placement of the CKUs in the text can be a marker for discourse structure in biological research articles.

3 Automatic detection of CKUs

According to our definition, a CKU is a factual proposition referring to a bio-event, and its discourse function is updating knowledge: its source is the author of the current article, and its basis is the experimental findings of the current

Title	Abstract	Introduction	Results	Figures	Discussion	Citation	Event representation
Interaction of survival of motor neuron (SMN) and HuD proteins [with m RNA cpg15rescues motor neuron axonal deficits]	Here **we** used mass spectrometry to **identify***HuD as* a novel *neuronal SMN-interacting partner*.	Here **we identify***HuD asa* novel *interacting partner of SMN,*	Together with our co-IP data, **these results indicate** that *SMN associates with HuD in motor neurons*	*SMN interacts with HuD.*	**Our** MS and co-IP **data demonstrate** a strong **interaction between SMN and HuD** in spinal motor neuron axons.	Furthermore, **these findings** are consistent with recent studies **demonstrating** that *the interaction of HuD with the spinal muscular atrophy (SMA) protein SMN* …	Entity1: HuD Entity2: SMN Relation: Interaction Location: Motor neurons

Table 2. The same bio-event repeated in the different sections of the paper, a citation, and its representation

article, and its basis is the experimental findings of the current article. The discourse function is indicated either by the proposition's structural position within the article or by metadiscourse.

We suggest detecting CKUs in three steps, combining state-of-the art document processing tools:

1. identifying structural discourse markers;
2. identifying rhetorical discourse markers,
3. extracting factual bio-events.

Structural indicators, i.e. the title, section titles or figure captions, are detected through markup in a straightforward way, if the article is encoded in a structured document format (e.g., XML). If this is not the case, a special conversion tool should be applied, as described in e.g. Déjean and Meunier (2007) to convert unstructured documents to structured documents.

Metadiscourse indicators, which convey both that the source of the new knowledge is attributed to the author(s) and that it is factual, such as "*here we demonstrate*", "*our results identify*", etc. could be detected by local pattern-matching rules in the majority of cases, since the authors often use highly recurring forms to express them. However, in some cases the expressions are somewhat more complex, and thus do not match local patterns. In order to ensure better performance, which is important due to the relevance and relatively small number of the claims to detect, we could apply the concept-matching methodology as described in Sándor (2007), which takes syntactic dependencies into account. This methodology consists of identifying specific kinds of metadiscourse as the realizations of patterns of concepts, which are present as semantic features in syntactically connected words and expressions.

To detect CKUs, we assume that these are indicated minimally by two co-occurring concepts: a first concept, which we call DEICTIC, and which conveys reference to the current work (*here*, *we, our, these*), and a second concept, which is a subclass of what we call MENTAL_OPERATION (*identify, demonstrate*, *find*, etc.). This specific subclass is a list of verbs and their nominalizations that belong to the category of "certainty verbs" in Thomas and Hawes (1994). This minimal pattern detects expressions like "*we identify*" or "*our finding*". In expressions like "*these results indicate*" or "*our data demonstrate*", the DEICTIC concept is linked to the certainty verb in an indirect way, since it is the modifier of the subject of the certainty verb.

This subject refers to the "base" factor of the bio-event (i.e. the indication comes from "results", and the demonstration from "data", see De Waard and Pander Maat (2009)), and thus it is also part of the metadiscourse. Its relevant semantic feature is called SCOPE in the concept-matching systems. In

summary, CKU-specific metadiscourse is covered by the pattern DEICTIC + SCOPE + MENTAL_OPERATION, where the "+" sign indicates a syntactic relationship.

Consider the three sentences containing CKUs in Table 1. The metadiscourse is in bold:

(1) **Here we** used mass spectrometry to **identify** HuD as a novel neuronal SMN-interacting partner.

(2) **Our analysis** of known HuD-associated mRNAs in neurons **identified** cpg15 mRNA as a highly abundant mRNA in HuDIPs compared with other known targets of HuD, such as GAP43 and Tau.

(3) Together with our co-IP data, **these results indicate** that SMN associates with HuDin motor neurons, and that these two proteins colocalize in granules within motor neuron axons.

While (3) follows a straightforward local pattern, in sentences (1) and (2) the relationship between "we" and "identify" and "our analysis" and "identify" needs deep syntactic analysis. This analysis is carried out by the Xerox Incremental Parser (XIP) (Aït et al. 2000), on top of which we have implemented concept-matching rules for detecting metadiscourse indicating CKUs.

We developed a simple concept-matching grammar based on the rules described above, and assessed the results of the automatic detection of the rhetorical indicators of CKUs in two papers. With respect to our manual annotation of CKUs the coverage is 81% and 80% and the precision is 62% and 51% respectively.

Once the metadiscourse is detected, another module should be applied for detecting bio-events, i.e. factual propositions that involve biological entities. This step can be executed by a state-of-the-art biological parser that detects factual bio-events, like the one by Nawaz et al. (2010). Subsequent integration of factual bio-event extraction should improve the precision, because the metadiscourse by itself does not guarantee the factuality of the bio-events, as in the following sentence:

(4) **Our findings provide** further support for the hypothesis that SMN can associate with multiple RBPs to regulate axonal mRNA levels in neurons, and that the different SMN–RBP complexes may be defined by their mRNA contents.

4 Validation: are CKUs indeed the main claims?

To test whether CKUs represent indeed the main claims of biology papers we carried out the following checks:

1. First, we asked a domain specialist both to validate the CKUs as main claims, and select them in two of full-text papers.
2. Second, we analyzed how a source paper is cited in other papers, and investigated whether the descriptions given in the referring texts correspond to the CKUs in the cited papers.

We discuss these forms of validation in turn.

4.1 Validation by domain specialists

We carried out the validation in two steps. In the first step we manually highlighted the CKUs in two papers according to the definition given in section 2, above, and asked a biologist to select the sentences that were relevant claims of the article. In this step all the CKUs have been validated. This indicates that if biologists are provided with a list of CKUs annotated by non-specialists based on discourse indicators, they do get access to relevant claims of the articles.

In the second step we asked the biologist to highlight the sentences that conform to the 6 points of our definition of CKUs. In the first article she selected 26 sentences, out of which only 12 sentences were conform to the definition of CKUs. The article contains 4 further CKUs, which the biologist did not select. Out of the 14 sentences that were highlighted by the biologist and that did not satisfy the definition of CKUs, 5 do not satisfy one important criterion of CKUs, that of factuality. The remaining 9 sentences were factual, but did not explicitly attribute the proposition to the authors of the article, i.e. did not contain metadiscourse that characterizes CKUs. In the second article the biologist selected 48 sentences, out of which 24 were indeed CKUs, and there is no more CKU is the article. Similarly to the first article, 3 out of the remaining sentences were not factual and 21 did not contain metadiscourse.

This experiment leads us to three interesting observations:

1. A list of CKUs is meaningful for the biologist, however, CKUs do not provide an exhaustive and well-definable list of main claims.

2. The definition of the CKUs is difficult to apply for a biologist who is not trained in rhetorical analysis.
3. The notion of a "main claim" is not straightforward to define formally.

4.2 Citing sentences collection

Work on citation-based summarization (e.g. Kaplan et al., 2009, Jbara and Radev, 2011, Nakov et al., 2004) focuses on creating 'a summation of multiple scholars' viewpoints […] using its set of citation sentences'. If we accept the premise of this work, which is that a collection of citation sentences offer a good overview of the cited papers, then CKUs should be well-represented in the collection of cited sentences. As a second check, we identified a collection of 20 citations of a full-text paper (Voorhoeve et al., 2006) and compared the citing sentences to the CKUs detected in this paper. We found that in all cases the citing sentences could be linked back to the CKUs (and indeed offer a good summary of the cited paper).

5 Discussion

5.1 Related work

De Waard and Pander Maat (2012) propose a model for epistemic classification of bio-events that consists of three parts: epistemic value (from factual through various degrees of certainty until lack of knowledge); base (grounding for the knowledge: reasoning, data or unidentified); source (author, named external source, implicit, attribution to the author, nameless external source, no source of knowledge). Each bio-event is characterized by a combination of the three factors. CKUs represent a special case in this system: their epistemic value is factual, their base is data derived from the work described in the article, and their source is the author. Whereas De Waard and Pander Maat do not differentiate among the various combinations of the factors, we propose to handle this unique combination on its own right, since it fulfills a special discourse function in the article, which facilitates access to the main claims.

Each of the three factors that characterize CKUs is taken into account in various text-mining systems, however, to our knowledge, no other system defines a discourse function in terms of these three factors. Nawaz et al. (2010) detect factual bio-events, but they do not detect authorship and base. The same holds for the annotation guidelines developed by and Wilbur et al. (2006). Teufel (2000) considers authorship but does not consider factuality and base. Blake (2010) differentiates among several kinds of base and considers only factual bio-events, but does not consider authorship.

Jaime-Sisó (2011) makes the same observation as we do: the authors summarize and repeat the main findings in every section of the articles. She attributes this phenomenon to the authors' adaptation to electronic publishing, where there is the possibility to navigate in the text. Repetition facilitates this navigation. Based on interviews with researchers and the analysis of 20 biology articles, she concludes that summarizing sentences that repeat the main findings in each section of biology articles are crucial both in writing and reading practices: "Aware of the scientists' reading practices, both editors and writers contribute to ensure that, whatever section of the text is scanned, and regardless of the reasons of approaching the article, the reader obtains the most newsworthy information, as if each of the sections could stand alone." (p. 87) "Noteworthy information" is mostly expressed by CKUs, although Jaime-Sisó does not provide a rhetorically based definition of summarizing sentences.

5.2 Proposed applications

We argue that the detection of Claimed Knowledge Updates constitutes a relevant goal for text-mining. CKUs are systematically signaled either by their position within the paper or by specific rhetorical discourse markers. This demonstrates that they constitute a systematic discourse organizing factor of articles. Moreover, CKUs can be detected by integrating state-of-the-art tools.

The detection of new factual knowledge could be useful in several tasks, such as summarization, information extraction, updating ontologies and knowledge bases, etc.

In particular, we wish to propose two use cases: first, the identification of CKUs could improve the output of automated knowledge bases that rely on text mining. Several text mining systems aim to provide multi-dimensional characterizations of bio-events, both academic systems such as

MEDIE[1]and iHoP[2], and commercial systems such as Ariadne[3]and BEL[4]. In none of these systems, however, are the various bio-events detected differentiated according to their role in updating knowledge. Showing only the CKUs, and not all the claims, would greatly enhance the efficiency and use of these automated knowledge bases. For example, the output of the query 'LATS2' as a subject in MEDIE returned the following sentences:

1. LATS2 is a member of the LATS tumor suppressor family.
2. The differences in the expression levels of the LATS2, S100A2 and hTERT genes in different types of NSCLC are significant.
3. LATS2 is a new member of the LATS tumour suppressor family.
4. Among the growing list of putative Mdm2-regulated proteins are several proteins playing a key role in the control of cell proliferation such as pRb, E2F1/DP1, Numb, Smads, Lats2 or IGF-1R.
5. In addition, modulation of novel target genes such as LATS2 and GREB1 were identified to be mediated by Nrf2.
6. Here, we show that LATS proteins (mammalian orthologs of Warts) interact directly with YAP in mammalian cells and that ectopic expression of LATS1, but not LATS2, effectively suppresses the YAP phenotypes
7. The tumor suppressor genes NEO1 and LATS2, and the estrogen receptor gene ESR1, all have binding sites for p53 and hsa-mir-372/373.

It is clear - even without studying the textual context - that not all of these sentences refer to a new finding pertaining to LATS2, which is what the user would like to see, and what a CKU parser would provide.

A second possible application of CKU detection could be the presentation of CKUs as metadata in biomedical publications, to aid the navigation within and among collections of biology articles. This is illustrated in a mock-up (Figure), which extends the PNAS publication scheme with an additional column presenting CKUs. The column in the middle is a part of the standard PNAS layout, and it points to the past, i.e. to existing articles that the current article draws on. But the third new column on the right extracts CKUs put forward in the current article. According to where the CKUs are, the readers can learn what type of arguments they could find to support them in the text to the left: in the introduction - background knowledge; in the results - experiments; in the discussion - various other links and implications; in the Figures - the illustration of the experiments.

To support both of these applications, CKUs could be marked up by the authors of the article during authoring or submission, making use of tools that identify CKUs. The systematic annotation of CKUs by the authors could provide them with a structural template against which they could check the article's coherence, and act in a role similar to a Structured Digital Abstract, proposed by Gerstein et al. (2007), as a 'computer-readable summary of pertinent facts'. These CKUs could then be added directly to a bio-event representation framework, where biological entities, interaction types, locations, etc. are structurally marked for easy information extraction. In this way, the user can easily track the grounding of a specific bio-event in past work, present experiments and future possibilities– and eventually, do better science.

[1]http://www.nactem.ac.uk/medie/
[2]http://www.ihop-net.org/UniPub/iHOP/
[3]http://www.ariadnegenomics.com/
[4] http://www.openbel.org

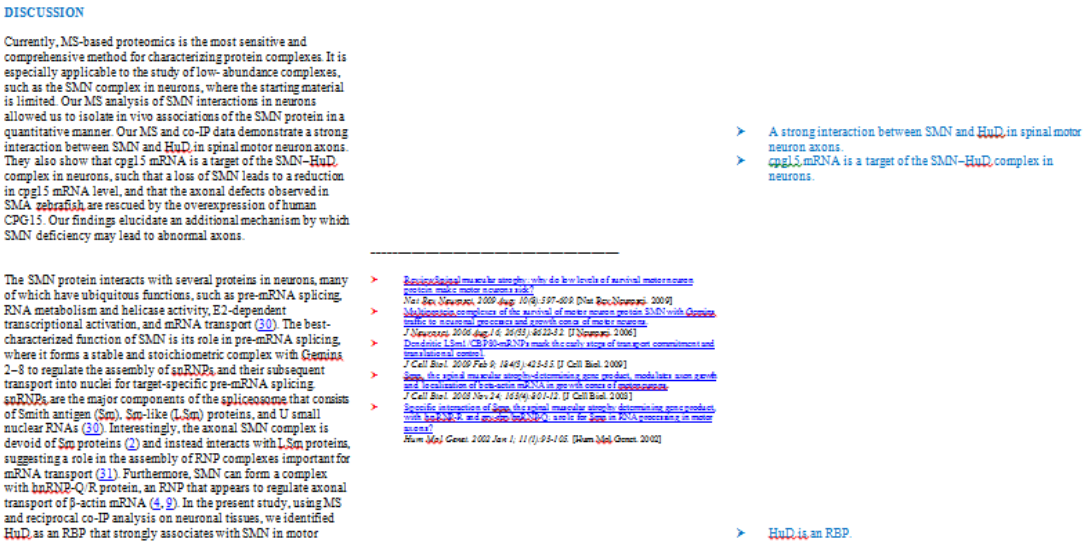

Figure Mockup of presenting CKUs in publications

Acknowledgements

We are indebted and deeply grateful to Prof. Maryanne Martone from the University of San Diego for her generous help in annotating the CKUs in our texts, and to our anonymous reviewers for helping us improve this paper.

References

Salah Aït-Mokhtar, Jean-Pierre Chanod, and Claude Roux. 2002. Robustness beyond shallowness: incremental dependency parsing. Natural Language Engineering, 8(2/3):121-144.

Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent Citation-Based Summarization of Scientific Papers. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 500–509, Portland, Oregon, June 19-24, 2011

Akten, Bikem, Min JeongKye, Le T. Hao, Mary H. Wertz, Sasha Singh, DuyuNie, Jia Huang, Tanuja T. Merianda, Jeffery L. Twiss, Christine E. Beattie, Judith A. J. Steen, and Mustafa Sahin. 2011. Interaction of survival of motor neuron (SMN) and HuD proteins with mRNA cpg15 rescues motor neuron axonal deficits, ProcNatlAcadSci U S A. 2011 Jun 21;108(25):10337-42.

Catherine Blake. 2010. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. Journal of Biomedical Informatics archive Volume 43 Issue 2, April, 2010

Pablo Ciccarese, Elizabeth Wu, June Kinoshita, Gwen Wong, Marco Ocana, Alan Ruttenberg, and Tim Clark. 2008. The SWAN Biomedical Discourse Ontology.J Biomed Inform. 2008 Oct;41(5):739-51. Epub 2008 May 4.. PMID: 18583197

HervéDejean and Jean-Luc Meunier. 2007. Logical Document conversion: combining functional and formal knowledge. Symposium on Document Engineering, Winnipeg, Canada, August 28-31, 2007.

Mark Gerstein, Michael Seringhausand and Stanley Field. 2007. Structured digital abstract makes text mining easy, *Nature* 447, 142 (10 May 2007) | doi:10.1038/447142a

Mercedes Jaime-Sisó. 2011. Summarizing Findings: An All-Pervasive Move In Open Access Biomedical Research Articles Involves Rephrasing Strategies. In Researching Specialized Languages.Studies in Corpus Linguistics 47.Edited by Bhatia, Vijay, Sánchez Hernández, Purificación and Pérez-Paredes, Pascual.Published by John Benjamins. Pp. 71-88.

Amjadabu Jbara and Dragomir R. Radev. 2011. Coherent citation-based summarization of scientific

papers. In Proceedings of ACL 2011, Portland, Oregon, 2011.

Dain Kaplan, Ryu Iida and Takenobu Tokunaga. 2009. Automatic Extraction of Citation Contexts for Research Paper Summarization: A Coreference-chain based Approach, Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP 2009, pages 88–95, Suntec, Singapore, 7 August 2009.

Maria Liakata, Simone Teufel, Advaith Siddharthan and Colin Batchelor. 2010. Corpora for conceptualisation and zoning of scientific papersProceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Malta.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction.International Journal of Medical Informatics. 75(6): 468-487.

Preslav I. Nakov, Ariel S. Schwartz, A., and Marti Hearst. 2004. Citances: Citation Sentences for Semantic Analysis of Bioscience Text, in the SIGIR'04 Workshop on Search and Discovery in Bioinformatics.

Raheel Nawaz, Paul Thompson, John McNaught, Sophia Ananiadou. 2010. Meta-Knowledge Annotation of Bio-Events. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010).

Cameron Neylon. 2012. Network Enabled Research: Maximise scale and connectivity, minimise friction, Blog post, February 2012, http://cameronneylon.net/blog/network-enabled-research/

Ágnes Sándor. 2007. Modeling metadiscourse conveying the author's rhetorical strategy in biomedical research abstracts. Revue Française de Linguistique Appliquée 200(2):97--109.

Simone Teufel. 1999. Argumentative Zoning: Information Extraction from ScientificText. PhD Thesis.

Simone Teufel and Marc Moens. 2000. What's yours and what's mine: Determining intellectual attribution in scientific text. Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora.

Sarah Thomas and Thomas P. Hawes. 1994. Reporting Verbs in Medical Journal Articles. English for Specific Purposes, v13 n2 p129-48 1994.

P. Mathijs Voorhoeve, Carlos le Sage, et. Al. 2006. A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors. *Cell*. 2006 Mar 24;124(6):1169-81.

Anita de Waard, Simon Buckingham Shum, Annamaria Carusi, Jack Park, Mathias Samwald, and Ágnes Sándor. 2009. Hypotheses, evidence and relationships: The HypER approach for representing scientific knowledge claims. In: Proceedings 8th International Semantic Web Conference, Workshop on Semantic Web Applications in Scientific Discourse. Lecture Notes in Computer Science, Springer Verlag: Berlin, 26 Oct 2009, Washington DC.

Anita de Waard and Henk Pander Maat 2009. Categorizing Epistemic Segment Types in Biology Research Articles. Workshop on Linguistic and Psycholinguistic Approaches to Text Structuring (LPTS 2009), September 21-23 2009

Anita de Waard,. and Pander Maat., H.P.M., 2012. Workshop on Detecting Structure in Scientific Discourse, ACL 2012, Jeju Island, Korea (this workshop).

W. John Wilbur, Andrey Rzhetsky and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction, BMC Bioinformatics, vol. 7, no. (356)