*Submission to Semantic Web Journal: Interoperability, Usability, Applicability*
*Special Issue on Survey Articles, Ed. Pascal Hitzler*

# Scientific Discourse on the Semantic Web: A Survey of Models and Enabling Technologies

**Simon Buckingham Shum**
Knowledge Media Institute, The Open University, UK *

**Tim Clark**
Massachusetts General Hospital & Harvard University, USA

**Anita de Waard**
Advanced Technology Group, Elsevier & Institute of Linguistics, Utrecht University, NL

**Tudor Groza & Siegfried Handschuh**
Digital Enterprise Research Institute, National University of Ireland, IE

**Ágnes Sándor**
Parsing & Semantics Group, Xerox Research Centre Europe, FR

**Abstract**

The desired outcome of all scientific endeavour is to advance the the body of accumulated knowledge in a materially verifiable way. This knowledge is communicated through the research literature, which presents scientific claims and their justifications through forms of discourse, expressed in document genres legitimated by a given research community. The study of the rhetorical and argumentative characteristics of such discourse has long-standing traditions, the results of which also provide insights into how scientific publishing, search and debate might take new forms on the social-semantic web. This article surveys, for a general readership, the growing body of work that models scientific discourse for social-semantic web applications, and offers a framework highlighting key features to help compare the various models. Secondly, we present examples of tools based on discourse models, which facilitate semantic navigation, structured debate, human and machine annotation of scientific texts, and literature analysis/alerting services. Finally, we identify some of the open research challenges confronting the field, and summarise the ways in which they are being tackled.

* Corresponding author: sbuckinghamshum@gmail.com / 0770 212 5734 / http://people.kmi.open.ac.uk/sbs

**Contents**

# 1 Introduction

This article introduces approaches that are being taken to answer the fundamental question: *What does scientific publishing and discourse look like on the social, semantic web?* We do not believe that the established scholarly article will disappear in the near future, since it is a very rich form of communication permitting many nuances, and it is culturally embedded in the literacies taught from childhood through higher education, not to mention its central role in the academic career trajectory. However, the scientific publication remains essentially a child of Gutenberg's printing press, and we now have a digital infrastructure unimaginable in the 15th Century. Beyond digital replicas of this paper artifact, our challenge is to invent a future in which the internet more radically improves the effectiveness in the ways in which we make, disseminate and contest knowledge level claims.

Where are we trying to get to? We are aiming to deliver sensible answers to queries which any scholar would consider fundamental to a critical perspective, for instance:

- What is the current state of the debate on this question?
- Who disagrees with this theory?
- Was this prediction ever fulfilled?
- What assumptions does this approach depend on?
- Are there different schools of thought around this problem?

Today's search engines and digital libraries provide little or no support for such queries, because they have little or no insight into the *discourse* that forms the heart of scholarly/scientific written communication, distinguishing it from other document corpora. The promising field of knowledge domain visualization (Shiffrin and Börner, 2004), based around scientometrics, is able to provide insight into the evolving structure of scholarly communities and terminological patterns within publications, but tells us little about the qualitative nature of that community's claims or debate. Claims to contribute to the literature in a given field are made using carefully constructed arguments, which vary between disciplines and their sub-communities. We teach our doctoral students how to construct knowledge-level claims in a manner which will bear peer review. The techniques and tools described in this article seek to formalize some of these patterns in order to deliver services and user interfaces which treat scientific knowledge not so much as a set of documents, but rather, as a network of meaningful, conversational moves which can be modelled as semantic relationships between nodes.

This survey is organised as follows: §2 reviews approaches to modelling scientific discourse; §3 reviews automated annotation of scientific discourse, particularly in publications; §4 introduces tools that deliver services based on these approaches, before §5 considers open challenges.

# 2 Approaches to modelling scientific discourse

## 2.1 Modelling the rhetorical structure of publications

### *2.1.1 Harmsze*

One of the first and probably the most comprehensive models for capturing the rhetoric and

argumentation within scientific publications was introduced by Harmsze [2000]. She focused on developing a modular representation for the creation and evaluation of scientific articles. Although the corpus used as a foundation for the analysis was about experimental molecular dynamics, the resulted model is uniformly valid for any scientific domain.

The author models the discourse by means of a coarse-grained structure split into modules and a series of links to connect these modules. The six modules proposed by Harmsze are as follows:

- **Meta-Information** is a support module that keeps the entire publication glued together. It consists of several parts, such as, the bibliographic information, abstract, lists of references or acknowledgements;

- **Positioning** sets the context of the research presented in the publication. It describes the *situation* in which the research issues are considered and the *central problem* of the research.

- **Methods** acts as a container for the authors' response to the central problem. The model provides three types of possible methods, i.e. experimental, numerical and theoretical methods.

- **Results** details the results achieved with the methods previously mentioned. It consists of *raw data* and the *treated results*.

- **Interpretation** contains the authors' interpretation of the results. It usually deals with the process of interpreting the results and the argumentation of the plausibility and on the relevance of the interpretation.

- **Outcome** aggregates the authors' *findings* and the *leads to further research*.

To connect the above mentioned modules, the model introduces two types of relations: (i) organizational links, and (ii) scientific discourse relations.

The *organizational links* provide the reader with the means to easily navigate between the modules composing the scientific publication. They connect only modules as entire entities and do not refer to the segments encapsulated in them, which in turn would identify the content. Harmsze distinguishes six types of organizational links: *hierarchical*, *proximity*, *range-based*, *administrative*, *sequential* and *representational*. On the other hand, regarding the links between segments of modules (scientific discourse relations), the model describes two main categories: *relations based on the communicative function*, that have the goal of increasing the reader's understanding and maybe acceptance of the publication's content, and *content relations*, that allow the structuring of the information flow within the publication's content. The first category is split into: *Elucidation*, as *Clarification* and *Explanation*, and *Argumentation*. The second category contains: *Dependency in the problem-solving process*, *Elaboration*, as *Resolution* and *Context*, *Similarity*, *Synthesis*, as *Generalization* and *Aggregation*, and *Causality*. Generally, the relations present an implicit polarity and don't have attached explicit weights or temporal aspects.

From the evaluation perspective, the authors performed a preliminary evaluation of the model, which showed that the model satisfies the purpose for what it was designed, but in reality, to our knowledge, it was not deployed in an actual application and consequently it failed to be adopted.

### *2.1.2 ABCDE*

A different discourse representation model was proposed by de Waard and Tel (2006). They identified a rhetorical block structure for scientific publications called ABCDE, similar to the IMRAD

(Introduction, Material and Methods, Results and Discussion)[1] structure. This proposal for conference proceedings in computer science was to develop LaTeX-stylesheet that identified five components in a document:

- **Annotations**, representing the set of shallow metadata associated with each publication (usually expressed in DublinCore[2] terms)
- **Background**, describing the positioning of the current research and the ongoing issues;
- **Contribution**, describing the work performed by the authors;
- **Discussion**, comparing the current work to other approaches, including implications and next steps;
- **Entities**, linked to references, personal names, project websites, etc.

The ABCDE proposal is that for each of the three content components (Background, Contribution and Dicussion), the author identifies a set of 'core sentences' (practically done within a LaTeX stylesheet). At rendering time, the LaTeX is converted into XML/XHTML and these core sentences can be contracted to form a (structured) abstract. For example, in a large conference, the Poster Session can be represented by displaying only the core Contribution sentences: the essence of what the authors did, and allow browsing to the Contribution section directly. The Entities include references and these are therefore explicitly not given at the end of the paper, but as triples that are referenced at the appropriate point. The vision is that the entity reference is a triple, with the referrong location in the citing document as the beginning, the DOI (Digital Object Identifier) of the referred document as the end point, and a rhetorical relationship such as the ScholOnto relations as the relationships. A basic evaluation of the ABCDE tools was performed on a collection of Semantic Wiki workshop papers.

Later work proposed to complement these structures with finer-grained annotation and a collection of relationship types, similar to Harmsze's (de Waard, 2008) but the model was too cumbersome to apply wholesale to a collection of papers. For an overview of these considerations, comparisons with Harmsze's work and to a modularly authored reference work, see (de Waard and Kircz, 2008).

The focus of this research then shifted to identifying linguistically definable segments (at the level of a clause) (de Waard and Pander Maat, 2009). In total, ten basic semantic segment types are defined: Fact, Hypothesis, Goal, Method, Result, Implication, Problem, Intertextual and Intratextual elements, and Regulatory elements (introducing other elements, of the type 'These results suggest that…'). There are clear linguistic clues for segments to belong to one or another category. Specifically, tense correspondences can be found between segments dealing with scientific concepts (Hypotheses, Problems, Facts and Implications), which are largely stated in the present tense, versus experiments (Goals, Methods and Results), largely stated in the past tense (de Waard and Pander Maat, 2009). Other defining features for segments are verb class (de Waard and Pander Maat, 2010), and a set of modality markers (not yet published). The main implementation possibilities for this work focus on the automated detection of rhetorical components in experimental discourse, and first attempt has to find these segment types computationally have been promising (de Waard, Buitelaar and Eigner, 2009).

---

[1] http://www.uio.no/studier/emner/hf/imk/MEVIT4725/h04/resources/imrad.xml

[2] http://dublincore.org/

Current collaborations are underway to unite this segment-centered view with work done on the meta-annotation of biological events (Nawaz et al, 2010) and the annotation of Core Scientific Concepts (CoreSCs) which are at the level of a sentence, and include Hypothesis, Motivation, Goal, Object, Background, Method, Experiment, Model, Observation, Result and Conclusion (Liakata et. Al, 2010).

### *2.1.3 SALT*

SALT (Semantically Annotated LaTeX)[3] (Groza et al., 2007a) is a semantic authoring framework targeting the enrichment of scientific publications with semantic metadata. SALT adopts elements from the Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) with the goal of modeling discourse knowledge items and their intrinsic coherence relations. The framework comprises three ontologies:

- **the Document Ontology** models the linear structure of a document, in terms of *Sections*, *Paragraphs* or *TextChunks*

- **the Rhetorical Ontology** captures the rhetorical and argumentation structure of the publication, and

- **the Annotation Ontology** connects the rhetorical structure present within the publication to its actual textual representation. This ontology forms a semantic bridge between the other ontologies, while at the same time capturing and exposing shallow metadata (title, authors, affiliations, etc.) and citation aspects, by re-using concepts from well-known vocabularies, such as DublinCore or FOAF[4].

The centre-piece of the framework is the Rhetorical Ontology that aims at decomposing the content of the publication into textual chunks having different granularity levels. This enables the creation of a coarse-grained rhetorical structure (similar to ABCDE – see previous section), and of a fine-grained semantic network, emerging from the connection of elementary discourse knowledge items. The latter is achieved via several types of relations that enable both the externalization of the rhetorical roles they carry, as well as the intrinsic argumentation, possibly spanning across multiple publications.

More concretely, the Rhetorical Ontology consists of three sides:

- **the rhetorical relations** side models the elementary discourse knowledge items, in terms of Claims and Supports at a higher semantic level, and Nuclei and Satellites (adopted from RST), at a lower linguistic level, together with a set of twelve rhetorical relations connecting them. The relations (e.g. A*ntithesis*, *Circumstance*, *Concession* or *Purpose*) are defined according to the original (Mann and Thompson, 1987) and extended[5] set of relations proposed by RST.

- **the rhetorical blocks** provide a coarse-grained structure for modeling the discourse, however, at a more detailed level than ABCDE. The list of rhetorical blocks include: *Abstract, Motivation, Scenario, Background, Contribution, Conclusion, Discussion, Evaluation, and Entities*.

---

3 http://salt.semanticauthoring.org/

4 http://www.foaf-project.org/

5 http://www.sfu.ca/rst/01intro/definitions.html

- **argumentation** side that captures the argumentation present in the publication via concepts like *Issue*, *Position* or *Argument*, adopted from the IBIS methodology (Kunz and Rittel, 1970).

Figure 1 depicts an example of SALT rhetorical structuring. The entire block is considered to be an Abstract rhetorical block, while at more fine-grained level, one can observe how the rhetorical elements are delimited, in addition to the rhetorical relations connecting them. The blue markup denotes a Claim ("*The solution relies in taking advantage ...*"), connected to the Support (the yellow markup – *"Consequently, the resulting knowledge becomes crystallized ..."*) by a Consequence rhetorical relation. The text contains also a Nucleus (*"From a technological perspective, ..."*) connected to two Satellites via the Means relations.

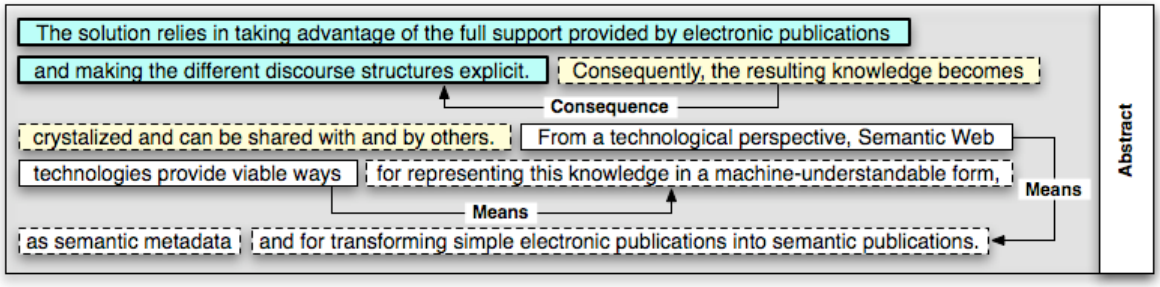



**Figure 1: Example of SALT rhetorical block and relations**

Using SALT on a scientific publication leads to a local instance model capturing the inter-connected linear, rhetorical and argumentation structures within that publication. At a later stage, the authors dealt with the global scope of modeling discourse knowledge items, i.e., items and relations that span across multiple publications. In (Groza et al., 2007b), following a semiotic approach (Ogden and Richards, 1923), the authors introduce a model for externalizing argumentative discourse networks.

## 2.2 Modelling argumentation

### *2.2.1 IBIS*

Much of the work on developing models for more expressing deliberation and argumentation in more structured ways traces back to the formative work of design and policy planning theorist Horst Rittel. Elsewhere (Buckingham Shum, 2003) we trace his work, whose characterisation in the 1970's of "wicked problems" has continued to resonate since: *Wicked and incorrigible [problems]...defy efforts to delineate their boundaries and to identify their causes, and thus to expose their problematic nature* (Rittel, 1972; Rittel and Webber, 1973). Rittel concluded that many problems confronting policy planners and designers were qualitatively different to those that could be solved by formal models or methodologies, classed as the 'first-generation' design methodologies. Instead, an *argumentative* approach to such problems was required: *First generation methods seem to start once all the truly difficult questions have been dealt with. …[Argumentative design] means that the statements are systematically challenged in order to expose them to the viewpoints of the different sides, and the structure of the process becomes one of alternating steps on the micro-level; that means the generation of solution specifications towards end statements, and subjecting them to discussion of their pros and cons*. (Rittel, 1972). We note that this call for more explicit, reflective discourse echoes the ideal forms of debate that research communications should display in the sciences (and for that matter, the kind of exploratory dialogue that has been found empirically to characterise quality learning conversations in diverse contexts (Mercer, 2004).

The resulting *Issue-Based Information Scheme (IBIS)* provided a set of conversational moves summarised in Figure 2, with the most commonly used highlighted in tools based on it such as gIBIS (Conklin and Begeman, 1988) and its direct descendant today, Compendium (Buckingham Shum, *et al.* 2006). The IBIS focus on making *Issues* (or Questions) explicit has been influential in shaping the ScholOnto and Cohere approaches described below, and the basic concept of linking arguments with *supports/challenges* semantics pervade most subsequent computer-supported argumentation

ontologies and tools.

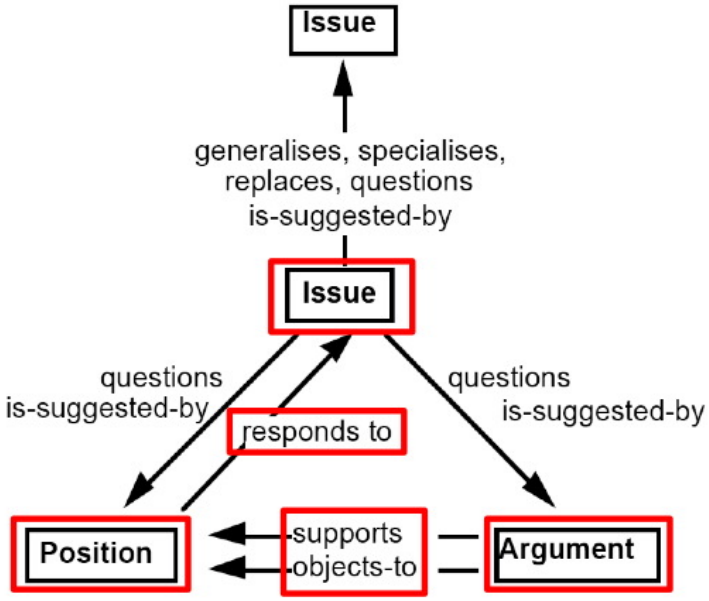



**Figure 2: The Issue-Based Information System**

### *2.2.2 ScholOnto*

Since a powerful interpersonal network sits behind the discourse network that is the literature, scientists take great care in the way in which they position their work in relation to their peers. In transitioning to an infrastructure with more explicit semantics, they must remain in control of the relationships they are seen to assert. Maintaining this social/semantic balance was a primary concern in the *Scholarly Ontologies (ScholOnto)* project[6], the first research programme to propose the decomposition of a scientific publication into elementary discourse knowledge items and meaningful relationships (Buckingham Shum, *et al.*, 1999; 2000; 2009). Building on small scale hypertext research systems from the late 1980s such as NoteCards (Halasz, *et al.*, 1987) and gIBIS (Conklin & Begeman, 1988), ScholOnto derived a taxonomy of discourse-centric relationships based initially on bottom-up empirical modelling of different literatures. The primary focus was not on modelling the rhetorical or linear structure of publications, but on providing ways for analysts to express the coherence relations *they perceive* between any of the following:

- extracts/paraphrases of document texts judged to be significant;
- the analyst's interpretations of texts — which may be very different from what the author actually wrote;
- "epistemic constructs" that are very real cognitively, but have yet to be grounded in any document — such as a problem, dilemma, hypothesis, prediction, etc. on which one is working.

Following hypertext terminology, all of the above serve as *nodes* whose contents are not machine processed, and whose granularity (i.e. size and complexity) is left up to the analyst using the tools. Nodes can be optionally classified to indicate the rhetorical role that the analyst wishes the node to

---

[6] Scholarly Ontologies Project: http://projects.kmi.open.ac.uk/scholonto

play in the network (e.g. Question, Argument, Evidence), or simply to highlight the status that the node has in a domain-specific model (e.g. Data, Language, Theory). The status and semantics of nodes are, however, of secondary interest (although language technologies and other semantic analysis tools could seek to align them with ontological entities).

The key focus is on the semantic relationships between nodes. In ScholOnto, two semantically connected nodes were said to constitute the making of a *Claim:* the analyst is asserting a meaningful relationships between two conceptual entities. In addition to nodes, the model allows for composite structures which correspond to the process of creating conceptual abstractions that cluster ideas into higher order constructs: claims can themselves be linked, and *sets* created to contain these and other sets.

In terms of relations, the proposed ScholOnto taxonomy comprised six types:

- **causal** links, e.g. predicts, envisages, causes or prevents;
- **problem related** links, e.g. addresses or solves;
- **similarity** links, e.g. is identical to, is similar to, or shares issues with;
- **general** links, e.g. is about, improves on, or impairs;
- **supports/challenges** links, e.g. proves, refutes, is evidence for, or agrees with;
- **taxonomic** links, e.g. part of, example of, or subclass of

Each relation had a positive or negative polarity, and a relative weight of 1 or 2. For example, the *causal* links *envisages* and *causes* have both a positive polarity, but different weights, the former being considered weaker that the latter. Similarly, *is inconsistent with* and *refutes* have a negative polarity with the latter being considered stronger than the former. In the subsequent move to a simpler Web 2.0 paradigm, this taxonomy was dropped in favour of a simple 3-way relational polarity: *positive, neutral* or *negative* links. Users are provided with some default discourse types under each heading, but are free to delete and extend the menu by inventing their own labels and tags on relationships, as dictated by their needs. All relationships are, however, positive, neutral or negative.

While the initial taxonomy was driven empirically "bottom-up", by modelling literatures from diverse fields spanning humanities, science and computing, later work explored the possibility of a theoretical basis for deriving and inter-relating discourse relations, based on psychology and computational linguistics research into *cognitive coherence relations* (CCR). Mancini and Buckingham Shum (2006) set out the progress made in this respect, hypothesising that it would be possible to define computational services based on structural patterns built from a CCR-inspired upper level relational ontology. Benn's doctoral work (2009; 2010) has reported progress in demonstrating the feasibility of this proposal.

Based on the above taxonomy, the ScholOnto Project prototyped a set of tools for the annotation and visualization of scientific argumentation (ClaiMaker, ClaiMapper, ClaimFinder) leading in subsequent projects to a Web 2.0 social-semantic tool (Cohere), whose design and evauluation are described later.

### *2.2.3 SWAN*

The SWAN (Semantic Web Applications in Neuromedicine)[7] project focuses on developing a semantically structured framework for representing biomedical discourse. Initially, the framework was applied to significant problems in Alzheimer Disease research, though it is not restricted only to this particular domain. The SWAN Ontology was developed to function as a the schema of a distributed knowledge base in Alzheimer Disease research, and to link information intrinsically captured with

[7] Semantic Web Applications in Neuromedicine: http://swan.mindinformatics.org/ontology.html

information residing in other biomedicine sources.

The basic idea behind SWAN is that scientific discourse at its core consists of presentation and discussion of a set of hypotheses, assertions or claims, and their supporting evidence. In presenting experimental results, a researcher must not only interpret the results, but place them in the context of other work. Ultimately a model of the phenomena being explored is presented, which may consist of a number of claims and various levels (from none to extensive) of supporting evidence. The claims made – particularly in complex phenomena such as observed in neurodegenerative disorders – may be consistent, or inconsistent, with those made by others in the field.

SWAN is a modular ontology which has endeavored to attain and keep alignment with other relevant terminology systems. Because it is modular, elements of SWAN may be used independently.

The SWAN ontology has the following modules, or sub-ontologies, which as noted can be used independently as needed:

- The SWAN Scientific Discourse Ontology provides the building blocks for defining the scientific discourse elements such as research statements (hypotheses, claims) and research questions. Supporting evidence for research statements is modeled using CiTO[8] [Shotton 2010] citations of FaBIO[9] bibliographic records – which in turn integrate the Dublin Core[10] and PRISM[11] publishing metadata specifications.

- The SWAN Scientific Discourse Relationships Ontology provides the sets of relationships for organizing the scientific discourse building blocks into a coherent story, and showing relationships between elements of different "stories", such as consistency / inconsistency.

- The SWAN Life Science Entities Ontology permits definition of such elements of biological concern as genes, proteins and organisms.

- The SWAN Agents Ontology permits specification of people, groups and organizations, for example, the authors of a publication, and integrates with FOAF[12].

- The SWAN Collections Ontology permits the creation of ordered lists, for example of authors.

- The SWAN Provenance, Authoring and Versioning Ontology declares and tracks the provenance of information declared in other SWAN modules.

- Tags, Qualifiers and Vocabularies supporting personal data organization, integrated with the SKOS[13] model of knowledge organization. [Miles and Bechhofer 2008].

- The SWAN Commons Ontology that provides the 'glue' to organize all the SWAN ontology

---

8 Citation Typing Ontology: http://purl.org/spar/cito/

9 FRBR-aligned Bibliographic Ontology: http://purl.org/spar/fabio/

10 Dublin Core Metadata Element Set, Verison 1.1: http://dublincore.org/documents/dces/

11 PRISM (Publishing Requirements for Industry Standard Metadata) Specification: Version 2.1: http://www.prismstandard.org/specifications/2.1/PRISM_prism_namespace_2.1.pdf

12 Friend of a Friend: http://www.foaf-project.org

13 Simple Knowledge Organization Schema: http://www.w3.org/2009/08/skos-reference/skos.rdf

modules into a coherent ontological framework[14].

SWAN has four types of discourse elements connected via a mixture of argumentation and cognitive coherent relations with implicit polarity (e.g. *consistentWith*, *inconsistentWith*, *discusses*, *alternativeTo*, *citesAsEvidence* or *inResponseTo*):

- **Discourse Element** is a narrative object, representing a mapping of digital resources to statements in natural language (e.g. sentences, paragraphs). Each such element may be linked dynamically to terms or statements in other domain ontologies;

- **Research Statement** represents a particular discourse element having a claim or hypothesis nature;

- **Research Question** is a topic under investigation;

- **Structured Comment** acts as a structure representation of a comment published in a digital resource.

Similar to the Scholarly Ontologies project, this model was also extensively evaluated and is currently in use by the Alzheimer Disease research community and others. An extended series of collaborations with other informatics and ontology researchers has since developed, centered in the HCLS Scientific Discourse Task of the World Wide Web Consortium (W3C). This has led to further integration of SWAN with the SIOC[15] ontology of blogs, wikis and discussion groups; the CiTO and FaBIO ontologies of citations and bibliographic records.[Passant, Ciccarese, Breslin and Clark, 2009].

Ongoing work in the W3C around SWAN is currently focused on the ideas of

- linking experimental design, conditions, materials and resultant data, plus required computations to process the results, to the claims they support;

- work with the textmining community to provide support for supervised algorithmic detection of claims and associated biomedical entities such as genes, proteins, metabolites, and biological processes, discussed in original text, to the formalized SWAN claims.

## 2.3 Comparison of discourse representation models by key features

In order to compare the preceding discourse representation models, the following set of features is used:

- **Course-grained rhetorical structure** – identifies the existence of a course-grained rhetorical structure representation within the model. Its goal is to capture the semantics of larger blocks of text inside the publication's content that have an associated rhetorical role.

- **Fine-grained rhetorical structure** – as opposed to the previous feature, this feature considers the fine-grained content composing the discourse (i.e. restricted discourse knowledge items in forms of claims, positions, arguments, etc) between which usually emerges a network arrangement driven by the different types of relations that connect the fine-grained elementary

[14] The SWAN Ontology Ecosystem: http://swan.mindinformatics.org/ontology.html

[15] Semantically-Interlinked Online Communities: http://rdfs.org/sioc/spec

items.

- **Relations** – looks at the types of relations used for linking the fine-grained structure into an unitary network.

- **Polarity** – specifies if the model includes explicitly the polarity of the relations (i.e. positive or negative). For example, a *supports* relation would have a positive polarity attached, while a *refutes* relation would have a negative polarity. Generally, this polarity is to some extent similar to the polarity extracted in the opinion mining and sentiment analysis field, which, we will not focus on, since it is out of the scope of this paper.

- **Weights** – specifies if the model considers explicitly the weights of the relations, i.e. if some relations are stronger than others. This feature can be tightly coupled to the polarity. For example, the *supports* relation might be considered stronger than the *agrees with* relation, both being positive from the polarity perspective.

- **Provenance** – indicates whether the model encapsulates also the provenance information attached to the fine-grained rhetorical structure (i.e. the accurate localization of the text span that represents the textual counterpart of the discourse knowledge item).

- **Shallow metadata support** – shows if the model has embedded support for shallow metadata (e.g. authors, titles, etc)

- **Domain knowledge** – analyses the close coupling of the model to particular domain knowledge areas.

- **Purpose** – presents the purpose, or intended use, of the model as envisioned by their creators.

- **Evaluation and uptake** – mentions the evaluation and uptake status of the model.

These last two features in the list try to capture the "practicality" dimension of the discourse representation models, with the last one pointing in essence to a reality-check, in terms of deployment, adoption and adequacy of the models in actual use by scientists.

Figure 3 presents a concise comparative overview of the five discourse representation models we have previously described. To make the first steps towards an unified discourse representation model, we believe that we have to find a proper balance between the features each of the currently existing models presents. In the following we will try to sculpt the skeleton of such an unified model, to be later discussed within the community.

<insert Figure 3 here>

**Figure 3: Comparative overview of the discourse representation approaches**

The first aspect to be considered is the overall structure of the model. By following a layered approach, such as the one proposed by SWAN and SALT, the unified model will gain flexibility, which in turn will be reflected in a more straightforward evolution. This would clearly decouple the rhetoric and argumentation from the provenance information, and from the shallow metadata and domain knowledge, while at the same time providing the opportunity for a modular enrichment of the model as a whole.

The second aspect is the discourse structuring level. To be able to capture the complete semantics hidden within the discourse, the model needs to address it at different levels. Consequently, it needs to

present both a coarse-grained structure, meaning the main sections of the paper. Current work under the aegis of the W3C Health Care and Scientific Discourse structure is aligning these efforts to basically correspond, for Life Science papers, to the IMRaD sections: Introduction, Methods, Results and Discussion. An OWL Ontology of Rhetorical Blocks is being discussed and compared with the Document Component Ontology . For papers in other fields, such as Computer Science, either the ABCDE or the SALT structure can be used, or a combination of both. At a more fine-grained level, key rhetorical elements of the paper can be represented, corresponding to ScholOnto's 'atomic nodes' or SWAN's 'discourse elements', possibly augmented with the segment types proposed in de Waard and Pander Maat (2009) (e.g., the key claims are Interpretation segments). Such a fine-grained representation allows the representation of the paper as a network of inter-linked elementary items that externalize the content's coherence and argumentation thread, connected via different types of rhetorical and argumentation relations, or 'Hypotheses, Evidence and Relationships' (de Waard, *et al.*, 2009).

Another remaining open question is the set of relations used to connect the elementary discourse knowledge items, as this is the point where the divergence between the existing models is the biggest. Having a closer look at the five sets of relations, we observe two distinct tendencies which can lead to a common denominator. On one side we have a mixture of cognitive coherent and argumentative relations (in the ScholOnto project, SWAN and Harmsze), while on the other side we have a more linguistic approach materialized in the rhetorical relations used by SALT. Both directions can be used in a complementary fashion. After a refinement of the rhetorical relations, we envision a co-existence of both sets, one modeling the argumentative support of the discourse, while the other capturing the coherence and rationale of the argumentation.

From the properties that relations can carry, we believe that *polarity* should be featured in the unified model, as it is extremely useful both for analysis and visualization of the discourse. The relations' weights are dependent on the extraction mechanism, and therefore should be defined by the corresponding approach and not included in the model, as such discrete quantifiers do not really provide a direct added value for an author / reader. The model also needs to contain the provenance information in addition to the shallow metadata describing the authorship and references.

Finally, the most important "non-functional" element to be considered when designing such an unified model is the adoption from the existing models of the lessons learned with regards to evaluation and uptake. The practical evaluation of the features to be selected for the model should play a crucial role in the overall design. Consequently, the resulting framework needs not only to be elegant and to satisfy all the requirements of a proper formal externalization, but also to be attractive for the average Web user. Contrarily, it will fail to achieve an appropriate community uptake and will remain just an elegant model on paper.

# 3 Machine annotation of scientific discourse

## 3.1 The challenge of automated annotation

In the past 50 years the largest amount of scientific discourse has been conveyed through journal and conference articles. Although other channels of scientific communication are fast evolving, as we have mentioned in the preceding sections, the article continues to be the standard academically accepted channel for transmitting research results.[16] Since the appearance of electronic publication of scientific

[16] Cf. "The number of scientific articles indexed by Thomson Reuters increased from fewer than 600,000 in 1990 to more than 1 million in 2009." (Times Higher Education online magazine: http://www.timeshighereducation.co.uk/story.asp?storyCode=412393&sectioncode=26)

articles huge efforts have been put into machine processing in order to researchers find their ways among the publications.

The main line of research aims at extracting factual information from the texts of the articles and transform them into structured data that can populate ontologies or databases (see e.g. . [Corney et al 2004, Garten et al. 2009]). Factual information extraction consists in extracting names and terms relevant for the domain and ontological relationships that hold among them. In the framework of factual information extraction each piece of extracted information is an entry in a flat data structure.

Scientific research, however, as we are arguing in this paper, does not consist in providing a list of facts, but it essentially consists in argumentation around facts. In the articles that account for their research the researchers make hypotheses, they support, refute, reconsider, confirm, build on previous ideas in order to support their own ideas and findings. Consequently the automatic processing of research articles should be able to capture and represent the evolution of ideas and findings, as they are described in the papers (for detailed argumentation see de Waard et al., 2009).

Research articles conform to rhetorical writing conventions that support the argumentative texture of the article and at the same time guide the reader in following it. The importance of these conventions, and thus the importance of rhetoric for composing research articles, is testified by the huge body and importance of literature describing the principles and techniques for writing research articles (See e.g. [Swales, 1990] and [Hyland, 2005] for a comprehensive picture). A recently evolving direction in natural language processing considers these rhetorical practices as the basis for extracting information embedded into the discursive, argumentative, rhetorical nature of the research articles. The knowledge items thus extracted are labelled according to their rhetorical status in the article: aim, result, conclusion, new knowledge, old knowledge, open question, etc. This labeling allows further processing in a nuanced way. Among other traditional applications like summary writing or information retrieval, automated rhetorical annotation can also assist curators to populate semantically structured knowledge bases like SWAN by pointing at hypotheses and claims, or can provide input to argumentative social network systems like Cohere by pointing at contested knowledge items.

There has been significant recent interest in shared formalisms for mapping elements of formal ontological structures to scientific or other documents on the web and in other formats such as PDFs, using models such as the Annotation Ontology[17] [Ciccarese et al 2010].

In order to illustrate the role of rhetorical development and argumentation in the constitution of knowledge conveyed in the research article, we present the first sentences of an abstract in biology. Factual information is in plain text, and rhetorically oriented expressions – often referred to as metadiscourse – are italicised.

> *Most evolutionists agree to consider that* our present RNA/DNA/protein world has originated from a simpler world in which RNA played both the role of catalyst and genetic material. *Recent findings from structural studies and comparative genomics now allow to get a clearer picture of* this transition. *These data suggest that* evolution occurred in several steps, first from an RNA to an RNA/protein world (defining two ages of the RNA world) and finally to the present world based on DNA. … (from: Forterre, 2005)

Discourse-oriented automated processing consists in the identification of the underlined elements and their interpretation in terms of rhetorical functions. This is a difficult task for two main reasons:

---

[17] Annotation Ontology (AO): http://code.google.com/p/annotation-ontology/

1. There exists a great variety of discourse and rhetorical models with various analysis units and goals.
2. It is notoriously difficult to map linguistic expressions into argumentative and rhetorical moves, since human languages do not provide special resources dedicated to rhetorical functions.

Owing to these reasons, the few existing computational linguistics applications to the rhetorical analysis of scientific articles[18] do not approach research articles through particular discourse linguistics models, but rather, propose robust discourse annotation methods inspired by a variety of models, while they rely on corpus analysis and are motivated by application needs.

In the following sub-sections we present and compare two approaches to the automated analysis of scientific articles: Argumentative Zoning and discourse analysis using a natural language dependency parser, the Xerox Incremental Parser. Among other work in this direction see e.g. [Pendar et al, 2008], [Ruch et. al 2007a], [Ruch et. Al 2007b], [Burstein et.al (2003)].

## 3.2 Argumentative Zoning

Argumentative Zoning was developed in the doctoral research of Simone Teufel (1999). This was the first attempt to automatically annotate rhetorical moves in research articles.

Teufel establishes the Rhetorical Document Profile (RDF), which is "designed to encode typical information needs of new readers in a systematic and structured way". As we see the emphasis here is on "information needs" and not on "information". The task is to automatically identify the parts of the articles that serve these information needs: what Teufel's calls the *argumentative zones*.

Argumentative zones, which cover the entire article, are defined in terms of a "model of prototypical scientific argumentation" containing the following argumentative moves:

| | |
|---|---|
| BACKGROUND | Generally accepted background knowledge |
| OTHER | Specific other work |
| OWN | Own work: method, results, future work. . . |
| AIM | Specific research goal |
| TEXTUAL | Textual section structure |
| CONTRAST | Contrast, comparison, weakness of other solution |
| BASIS | Other work provides basis for own work |

This list is inspired and motivated by a variety of approaches to the analysis of scientific discourse. From a discourse analysis point of view it draws on Swales' model of argumentative moves [Swales] and it uses Hyland's system of the description of metadiscourse [Hyland 1998]. From a practical point of view it aims at fulfilling requirements for detecting the attribution of intellectual ownership, citations and author stance. It also applies work on problem solving processes (e.g. [Hoey,1979], [Solov'ev]), and on the strategies of scientific argumentation [Swales, 1990], [Kircz, 1998].

Argumentative zoning is described as a difficult task both from the point of view of the establishment of a gold standard for annotation and from the point of view of automated execution. Teufel concludes

[18] There exists a great body of work in computational discourse analysis (cf. [Marcu (2000)], [Mann-Thompson (1987)], [Polanyi et al. (2004)]), however their categories and methods have not been applied to robust processing, which is required in information extraction tasks.

that new evaluation methods are required, since the interpretation of the results in terms of recall and precision is not straightforward either.

Originally, argumentative zoning was proposed for automatic summarization and information retrieval tasks. Later it was also used for educational purposes [Feltrim et al 2005] and citation indexing [Teufel 2005]. Since the theory and technique of argumentative zoning are shown to be robust and operational, subsequent work consists in annotation experiments in different disciplines, like chemistry [Teufel et al. 2009] and biology [Mizuta et al. 2006].

### 3.3 Using the Xerox Incremental Parser for detecting salient sentences

Sharing the basic assumption of argumentative zoning, i.e. that rhetorical moves can be detected from the author's language use, a different approach is taken in several applications developed with the Xerox Incremental Parser (XIP) [Aït, et al. 2002] for the rhetorical analysis of scientific articles. Instead of covering the whole article, this approach aims at highlighting the main research issues that the articles handle.

XIP annotates the following rhetorical functions as bearing the main research issues:

| | |
|---|---|
| SUMMARIZING | summarizing aims, claims, results, conclusions |
| BACKGROUND KNOWLEDGE | descriptions of previous ideas |
| CONRASTING IDEAS | descriptions of ideas as contrasting |
| NOVELTY | descriptions of new ideas |
| SIGNIFICANCE | descriptions of ideas as being significant |
| SURPRISE | descriptions of ideas as being surprising |
| OPEN QUESTION | descriptions of open questions |
| GENERALIZING | descriptions of research trends |

The choice of the rhetorical moves annotated as bearing the main research issues is motivated by various considerations. SUMMARIZING and BACKGROUND KNOWLEDGE relate to conveying main ideas in a straightforward way in the rhetorical construction of research articles. The other categories have their roots in Thomas Kuhn's view of science as primarily a problem-solving activity [Kuhn 1962]. Thus the *raison d'être* of any research paper is the problem, and the main ideas are to be found in sentences where the research issues are described. These sentences fulfil rhetorical functions of contesting, questioning or emphasizing research-related ideas, facts or theories as being significant or new research-related ideas, facts, or theories, of indicating a gap in knowledge, or of pointing out any flaw or contrast related to the research topic. This approach does not claim to provide a complete characterization of the research problem, neither does it represent the rhetorical construction of the article, but its main goal is to provide assistance in rapidly gaining understanding about the approach of the article to the research in question.

The rhetorical functions detected by XIP partly overlap with the argumentative zones, and partly are different from them. The main difference is that the contrasts among ideas are not approached from the point of view of intellectual ownership, but rather from the point of view of the various ways in which contrasting ideas are introduced.

There have been a number of proof-of-concept applications that justify the choice of these categories as bearing salient ideas:

1. Detecting abstracts in the Pubmed database that describe substantially new findings [Lisacek et al, 2005]
2. Improving information retrieval in a search engine dedicated to educational science [Sándor and Vorndran, 2010]
3. Reading assistance for peer-reviewers [Sándor and Vorndran, 2009]

4. An ongoing experiment for extracting research issues in project reports (Buckingham Shum, *et al.* 2010)

# 4 Examples of semantic tools for scientific discourse

## 4.1 Tools for human authoring & annotation

### *4.1.1 ClaiMaker & ClaiMapper*

*ClaiMaker* was the first web application developed in the ScholOnto project, providing a user interface for building semantic hypertext networks using the scholarly discourse relations taxonomy described earlier (Figure 4). Being developed in 2001-04, it was "pre-Web 2.0" in capabilities and user experience, but served as an early research prototype to investigate usability, modelling and system development issues (Buckingham Shum, *et al.* 2007).

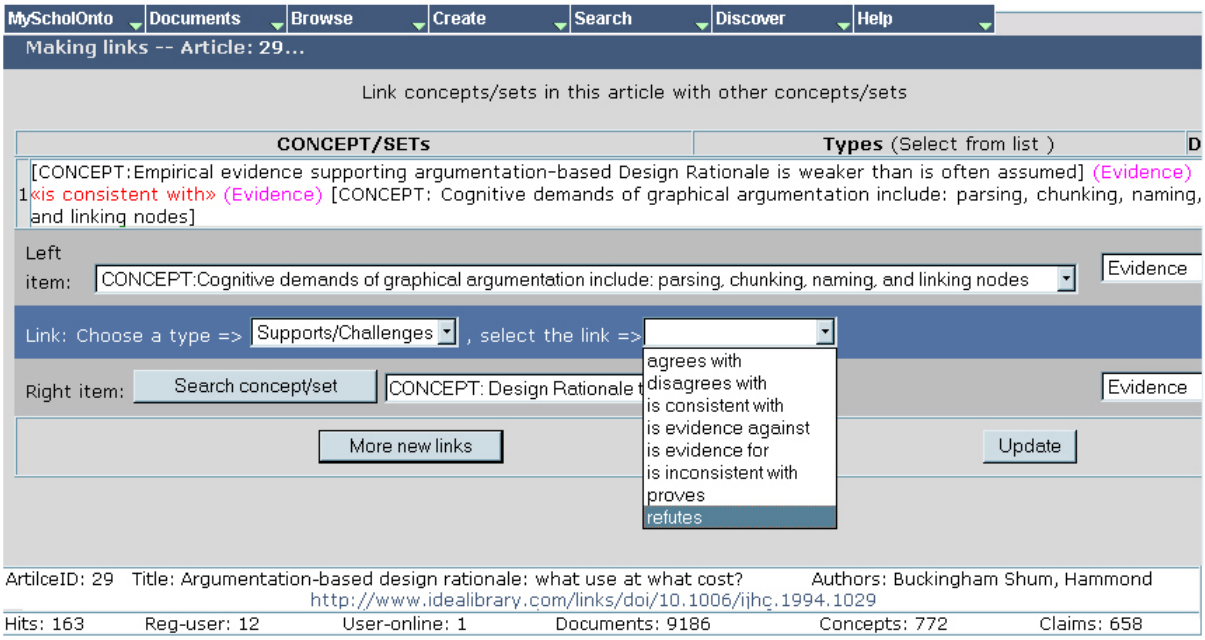



**Figure 4: The ClaiMaker forms interface for creating a claim. The bottom bar gives details of the paper the reader is modelling. The user has already selected the concept to be linked from and given it the optional type "Evidence". She is currently selecting a link from the drop down list of options. The next step will be to select the search button to look for the third component of the Claim triple (Buckingham Shum, et al. 2007).**

*ClaiMapper* was a visual hypermedia tool sketch rough maps of the literature using ClaiMaker's scheme, representing a *claim* as a semantic triple, which could itself be linked to as a composite node, to build chains and more complex structures (Figure 5).

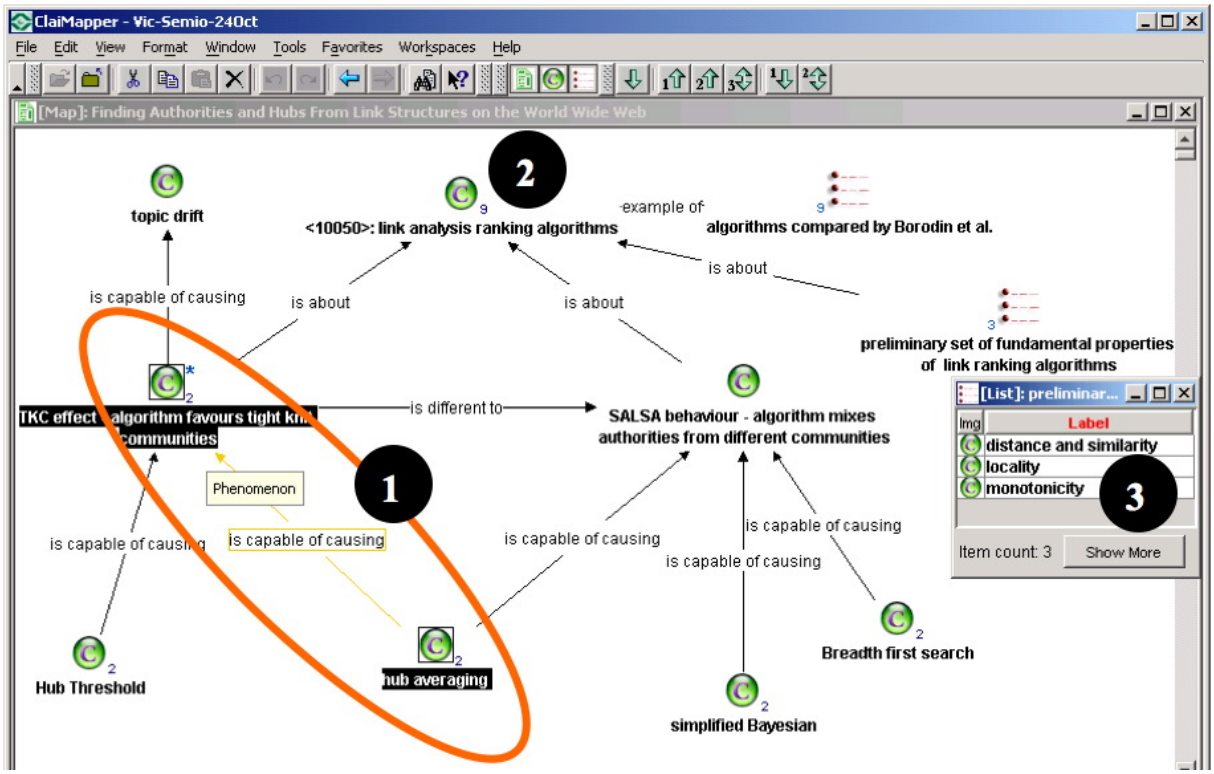



**Figure 5: Sketching ClaiMaker compatible models using the ClaiMapper tool. (1) In the circled Claim, the node *TKC effect…* has the type (i.e. plays the role of) Phenomenon. (2) The Concept *link analysis ranking algorithms* is shown as being used in 9 different contexts. (3) On the right is a Set named *preliminary set of fundamental properties of link ranking algorithms*, which when opened lists three concepts which the analyst has found. (Buckingham Shum, et al. 2007).**

*Cohere*[19] was released in 2008 (Buckingham Shum, 2008), as a development of the 2001-04 era ClaiMaker tool, taking advantage of Web 2.0 functionality such as social networking, highly interactive web user interfaces, open source browser extensions, snippet code to embed nodes and maps in external websites, syndication feeds, and RESTful APIs to loosely connect independent services via standards-based data formats such as RDF. In particular, it provides direct web annotation capability via a Mozilla Firefox browser extension, in order to anchor semantic annotations (*problem; hypothesis; assumption, etc.*) in any source web document (Figure 6).

[19] Cohere: http://cohere.open.ac.uk


**Figure 6: Annotation of a website using Cohere's Firefox extension sidebar (De Liddo and Buckingham Shum, 2010). These nodes can then be semantically connected to others.**

### *4.1.2 SALT LaTeX syntax and Word plugin*

Manual Semantic Authoring process refers to the process of manually enriching a scientific publication with explicit linear, rhetorical and argumentation structures, while authoring the publication. This process is conceptually independent of the authoring environment, however from an implementation point of view it differs with respect to whether the authoring is done in LaTeX or MS Word. In the context of SALT, as a proof of concept, the authors have developed Semantic Authoring mechanisms for both:

- LaTeX, in order to support scientific communities such as Physics, Mathematics or Computer Science, where it is considered to be the *de facto* standard for scientific authoring and publishing , as well as for

- MS Word (2003), to support other communities, e.g., Biomedical.

LaTeX is a high-quality typesetting system that enables the authoring of documents in a programmatic manner. Instead of following a visual (component-driven) approach like MS Word, LaTeX introduces a series of commands that the writer uses to produce the formatting and style of the text. The LaTeX author will be familiar and comfortable with this programmatic approach when writing the actual content of the publication.

LaTeX provides the most natural environment for Semantic Authoring. While using other writing environments, the annotation would impose a serious overhead, due to its characteristic "programming" feature, in LaTeX, this overhead, although still existing, is reduced to the minimum. In addition, the LATEX author has a special mindset, as one could generalize that the LaTeX commands represent themselves an annotation for the content. These reasons led to the development

of a set of special commands to facilitate SALT annotations.

- The authors introduced corresponding elements for each of the three sides of the SALT Rhetorical Ontology:Rhetorical blocks, i.e., chunks of text with a length carrying from a few sentences to one paragraph, are defined as LaTeX environments (e.g., *\begin{motivation} ... \end{motivation}*),

- Elementary discourse knowledge items (text chunks with a smaller length) together with their rhetorical relations are defined using LaTeX commands, similar to the style or formatting ones (e.g., \claim[ID]{ ... }, \cause{CLAIM_ID:SUPPORT_ID}) – see Figure 7

- Argumentation elements, being also elementary discourse knowledge items, are defined as well via LaTeX commands (e.g. \position[ID][CLAIM_ID]{ ... }).

One particularity of the elementary discourse knowledge items commands is the presence of identification elements (i.e., ID, or CLAIM_ID), required to enable the modeling (and explicit creation) of the rhetorical relations existing between them. These need to be created and managed by the author, since LaTeX does not provide any means for tracking the localization, within the content, of a particular chunk of text.

```
One remark that needs to be made at this point is that \claim[c1]{we have not defined
commands for nuclei and satellites}, \support[s1]{because of the high complexity of the
annotation task.} \cause[r1]{c1:s1} While in the case of claims and supports one can prove
the benefit of their externalization, for the other two concepts, although there is a clear benefit,
this is hard to grasp by the usual author. We believe that \support[s2]{if an author would
become familiar with marking up all the nuclei and satellites and the rhetorical relations,}
\claim[c2]{it would help her structure much easier her thoughts and argumentation and thus
lay down the publication's "red thread" in a straightforward manner.} \condition[r2]{c2:s2}
Unfortunately such a familiarization process has a very high learning curve, mostly dependent
on the psychological and knowledge background, as well as on the structuring mechanism of
this cognitive task.
```

**Figure 7: Example of SALT LaTeX annotation**

Unlike the previously described LATEX environment, MS Word follows a strictly visual, component-based paradigm. The author of a publication is focused only on writing the actual content, while the formatting and style is mostly done by the environment with small hints from the writer by means of some "clicks". In this case, the Semantic Authoring process imposes a higher overhead, because one needs to select the piece of text to be annotated, and then activate the actual markup, via a contextual menu or via options provided in the application's bar menu. Such a procedure usually defocuses the author, as she needs to switch from a writing mindset to an annotation mindset. Nevertheless, from a conceptual point of view, interleaving the writing and annotation processes is possible and thus MS Word can enable Semantic Authoring.

To cover this shift in paradigm, the authors developed an MS Word module (similar to ClaiMaker) that enables the writer to enrich publications with SALT instances. This provides the necessary visual components to create the markup, in forms of a contextual menu, a side panel and an associated bar menu. Figure 8 depicts an example of SALT annotation of elementary discourse knowledge items using MS Word.

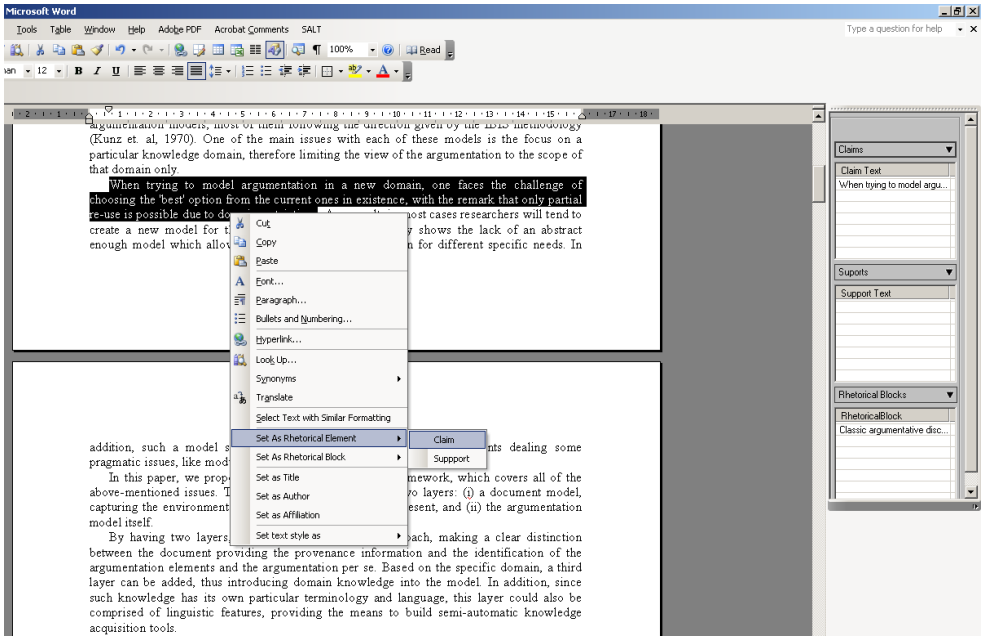



**Figure 8. Example of SALT MS Word annotation**

### *4.1.3 The SWAN Workbench*

The SWAN Workbench was designed to support the construction of the AlzSWAN knowledge base[20]and allowed a knowledge base curator to model a biomedical hypothesis, claims, evidence, discourse releationships and associated bio-entities (genes, proteins, organisms). The persistence mechanism is a triple store – literally everything is modeled as RDF triples. While it allowed the initial construction of AlzSWAN, the process was relatively labor-intensive. This product is now being replaced by the SWAN Annotation Framework (see below), in which the annotator receives extensive support from textmining algorithms, while being able to override the decision of the algorithms and/or to include pure manual annotation.

## 4.2 Tools for automated annotation

### *4.2.1 XIP*

Based on the approach introduced in §3.3, XIP annotates salient messages in research articles and labels them according to a list of rhetorical functions. In order to illustrate the outcome of this analysis we present here sample results.

Figure 9 is an interface for searching Pubmed abstracts. It calls an algorithm that ranks higher abstracts where sentences labelled highlighted in dark blue are present [Lisacek et al. 2005]. These are sentences that describe both “CONTRASTING IDEAS” and “NOVELTY” (see 3.3. above). The sentences highlighted in light blue describe any of the other rhetorical functions, and the ones in yellow fulfil the “SUMMARY” function. The query words and automatically detected related words are coloured.

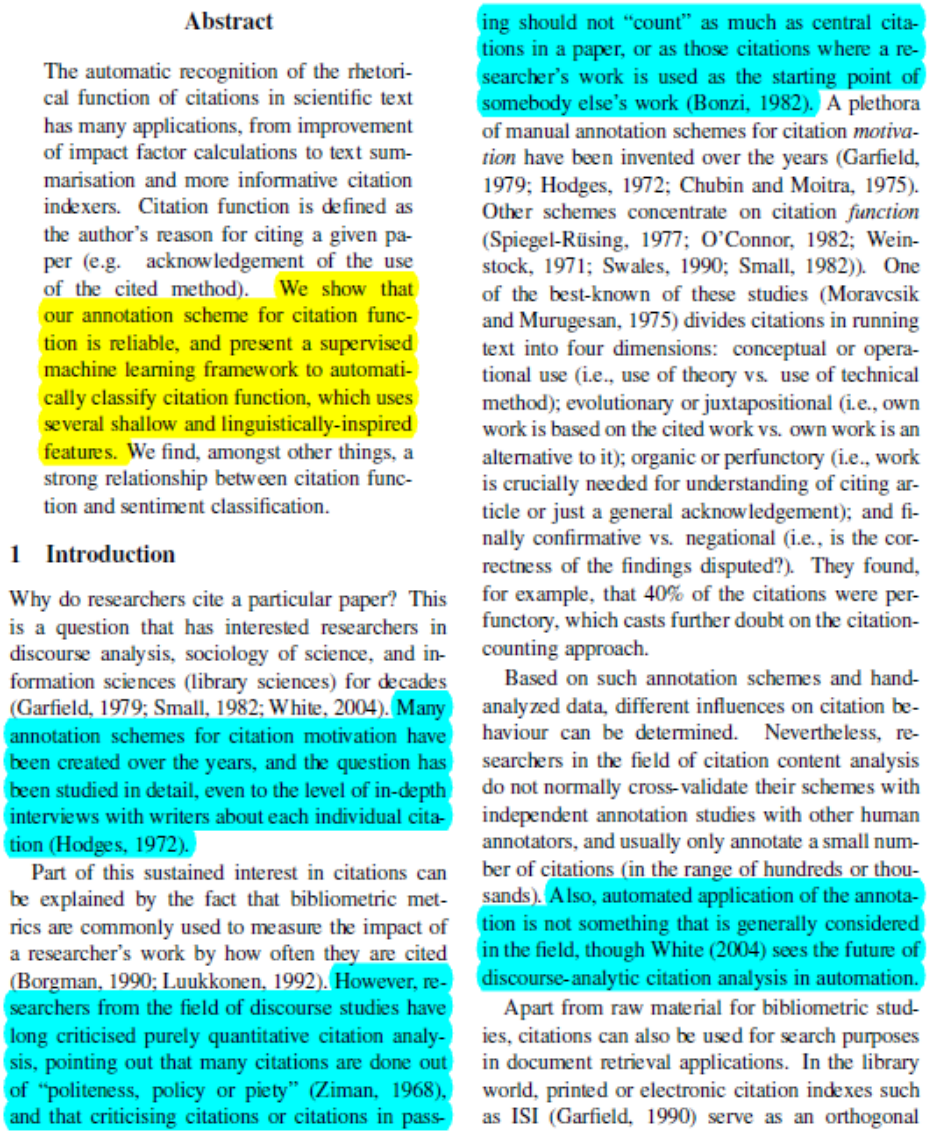
[20] AlzSWAN: http://hypothesis.alzforum.org/swan/


**Figure 9: Example output from XIP, highlighting passages in Pubmed abstracts identified as potentially significant based on their rhetorical status**

Figure 10 is a page of an article in the domain of computational linguistics. The sentences highlighted in yellow are the ones labelled “SUMMARY”, and the sentences highlighted in blue are the one’s that get any of the other labels.

An expert user recognizes the metadiscourse conveying the rhetorical functions in the highlighted sentences as well as their categories (e.g. in Figure 9 CONTRASTING IDEA: “… has been reported … but … remains controversial”, in Figure 10 BACKGROUND KNOWLEDGE: “… generally considered in the field”.), whereas, readers would not spontaneously and rigorously label the sentences as the machine does. This is supposedly due to the way people read articles: despite their importance in the argumentative development of the article, rhetorical functions are not explicitly recognized during the reading process, as they are not explicitly expressed by language either. By rendering explicit such underlying relevant aspects as rhetorical functions, the machine contributes to the sense-making process as it is supported by the proof-of concept experiments mentioned in section 3.3.

Future plans for development include the enrichment of the annotation categories, and thus providing a more complete list of the rhetorical functions that serve the description of research problems; studying annotations across various research fields (natural vs. social sciences); detecting the domain-specific concepts described in the salient sentences, and thus providing the rhetorically supported list of key concepts of the articles; and finally finding new applications where this kind of rhetorical annotation is beneficial.



**Abstract**

The automatic recognition of the rhetorical function of citations in scientific text has many applications, from improvement of impact factor calculations to text summarisation and more informative citation indexers. Citation function is defined as the author's reason for citing a given paper (e.g. acknowledgement of the use of the cited method). We show that our annotation scheme for citation function is reliable, and present a supervised machine learning framework to automatically classify citation function, which uses several shallow and linguistically-inspired features. We find, amongst other things, a strong relationship between citation function and sentiment classification.

**1 Introduction**

Why do researchers cite a particular paper? This is a question that has interested researchers in discourse analysis, sociology of science, and information sciences (library sciences) for decades (Garfield, 1979; Small, 1982; White, 2004). Many annotation schemes for citation motivation have been created over the years, and the question has been studied in detail, even to the level of in-depth interviews with writers about each individual citation (Hodges, 1972).

Part of this sustained interest in citations can be explained by the fact that bibliometric metrics are commonly used to measure the impact of a researcher's work by how often they are cited (Borgman, 1990; Luukkonen, 1992). However, researchers from the field of discourse studies have long criticised purely quantitative citation analysis, pointing out that many citations are done out of "politeness, policy or piety" (Ziman, 1968), and that criticising citations or citations in passing should not "count" as much as central citations in a paper, or as those citations where a researcher's work is used as the starting point of somebody else's work (Bonzi, 1982). A plethora of manual annotation schemes for citation *motivation* have been invented over the years (Garfield, 1979; Hodges, 1972; Chubin and Moitra, 1975). Other schemes concentrate on citation *function* (Spiegel-Rüsing, 1977; O'Connor, 1982; Weinstock, 1971; Swales, 1990; Small, 1982)). One of the best-known of these studies (Moravcsik and Murugesan, 1975) divides citations in running text into four dimensions: conceptual or operational use (i.e., use of theory vs. use of technical method); evolutionary or juxtapositional (i.e., own work is based on the cited work vs. own work is an alternative to it); organic or perfunctory (i.e., work is crucially needed for understanding of citing article or just a general acknowledgement); and finally confirmative vs. negational (i.e., is the correctness of the findings disputed?). They found, for example, that 40% of the citations were perfunctory, which casts further doubt on the citation-counting approach.

Based on such annotation schemes and hand-analyzed data, different influences on citation behaviour can be determined. Nevertheless, researchers in the field of citation content analysis do not normally cross-validate their schemes with independent annotation studies with other human annotators, and usually only annotate a small number of citations (in the range of hundreds or thousands). Also, automated application of the annotation is not something that is generally considered in the field, though White (2004) sees the future of discourse-analytic citation analysis in automation.

Apart from raw material for bibliometric studies, citations can also be used for search purposes in document retrieval applications. In the library world, printed or electronic citation indexes such as ISI (Garfield, 1990) serve as an orthogonal

**Figure 10: Example XIP output, annotating a computational inguistics article**

### *4.2.2 SWAN Annotation Framework*

The SWAN Annotation Framework (AF) was developed in collaboration with a major U.S.-based pharmaceutical company and in conjunction with the NIH-supported Neuroscience Information Framework (NIF), It is currently in alpha release to project participants, with anticipated production release in 2011 as part of the NIF. [Ciccarese et al, in preparation]

AF is a three-tier application which associates the URIs of selected ontological elements (terms) with localized parts of a web document, using the Annotation Ontology as its information schema. The Client tier is written in Javascript and may be embedded in any web application by inclusion of a few

lines of code. The Middle tier is written using the open source GRAILS[21] software framework [Rocher, G and Brown, J. 2009]. It provides access by the client to textmining functionality via web service calls, and to local persistence services via database calls, while proxying the selected web document to the local client. The Persistence Layer is implemented in mySQL using a schema mapped to AO. The current texmining service offering is the NCBO Annotator Service [Shah et el. 2009], with other services envisioned by production release. All annotation in AF is stand-off, i.e., fully decoupled from the document being annotated, and fully provenanced. [Ciccarese et al, in preparation]

SWAN Annotation Framework is notable in that its ontology is not SWAN, but AO, which is intentionally orthogonal to any particular domain ontology. This means that any ontology of document structure, and any terminology system can in principle be applied as annotation to a web document using the AF. [Ciccarese et al, in preparation]

## 4.3 Tools for browsing and searching

### *4.3.1 ClaimFinder & Cohere*

ClaimFinder was a research prototype which delivered the search services based on data authored in ClaiMaker (see above). The default page provided a simple, single-field form for users to do keyword searching, with 'advanced' search tabs delivering encapsulated services such as *Perspective Analysis* and *Lineage* (Buckingham Shum, et al. 2007). On invoking a ClaimFinder service, the tool generated interactive visualizations of the argumentative claim structures in which the relevant Concepts/Sets/Claims were embedded (e.g. Figure 11). These could be browsed by selecting a node to see the underlying detail, the source document it originated from, or to reveal/hide structure by zooming, rotating or filtering the number of links from the selected node.

[21] GRAILS: http://www.grails.org/

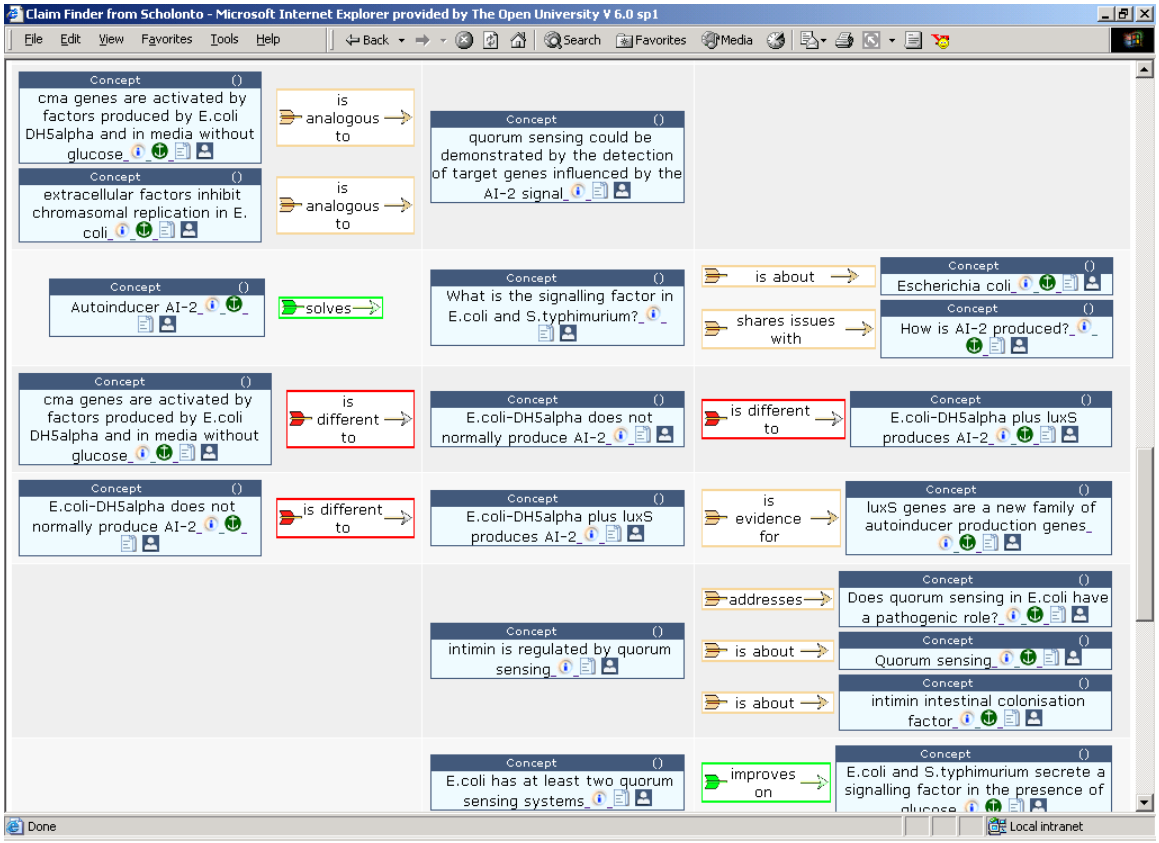



**Figure 11: ClaimFinder generates interactive visualizations of argument structures in response to queries. In this rendering, a three-column tabular layout shows each Concept/Set in the search results, with incoming and outgoing links to Concepts/Sets in the left and right columns.**

Cohere (introduced above) generates lists of nodes, websites, and visualizations of the semantic hypertext networks in its database (De Liddo and Buckingham Shum, 2010). As an example of how it helps users to exploit semantic connections, 'agents' can be set to watch the network for contributions that match semantic connections of interest to the user (Figure 12).

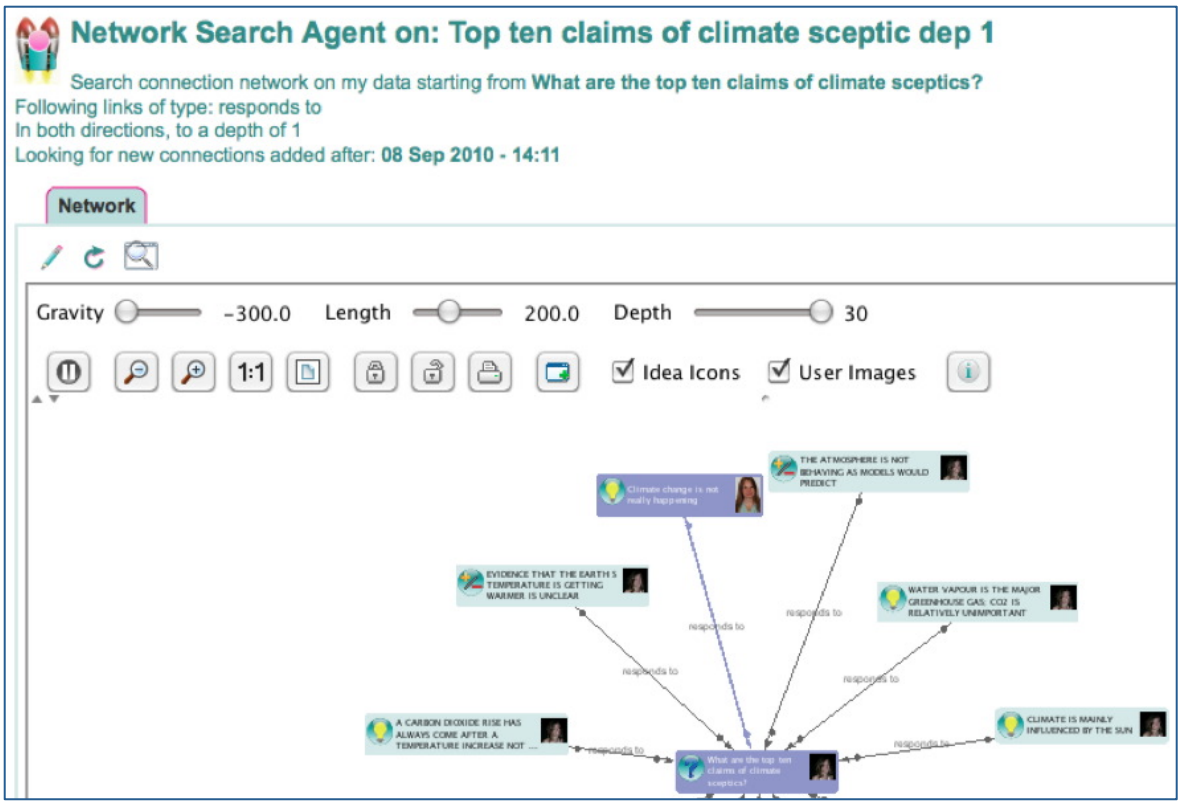



**Figure 12: An agent set to watch the network for connection types of interest, emails the user with an alert link, which when followed generates a map highlighting relevant connections (Buckingham Shum and De Liddo, 2010).**

### *4.3.2 KonneX$^{SALT}$*

As described in the previous sections, SALT lays the foundation for modelling the rhetorical and argumentative structures within scientific publications, while its associated LaTeX syntax and Word plugin, enable the enriching of these publications with semantic annotations. In order to full advantage of the resulting semantic metadata, (Groza et al., 2008) have developed KonneX$^{SALT}$, a claim federation infrastructure.

This had two main goals: (i) to provide a unique and global access point to the SALT semantic metadata, created via different means and published via KonneX$^{SALT}$, and (ii) to act as a basic lookup service for explicit knowledge. However, the main value provided by the tool was the capability of browsing argumentative discourse networks that span over multiple publications. As shown in Figure 13, starting from a discovered claim, the user was able to visualize not only its different textual representations (possibly co-existing in several publications), but also the argumentation it provides for positions referring to claim present in other publications. In this way, KonneX$^{SALT}$ provides the necessary means for the user to weave the web of claims hidden within the content of these scientific publications.

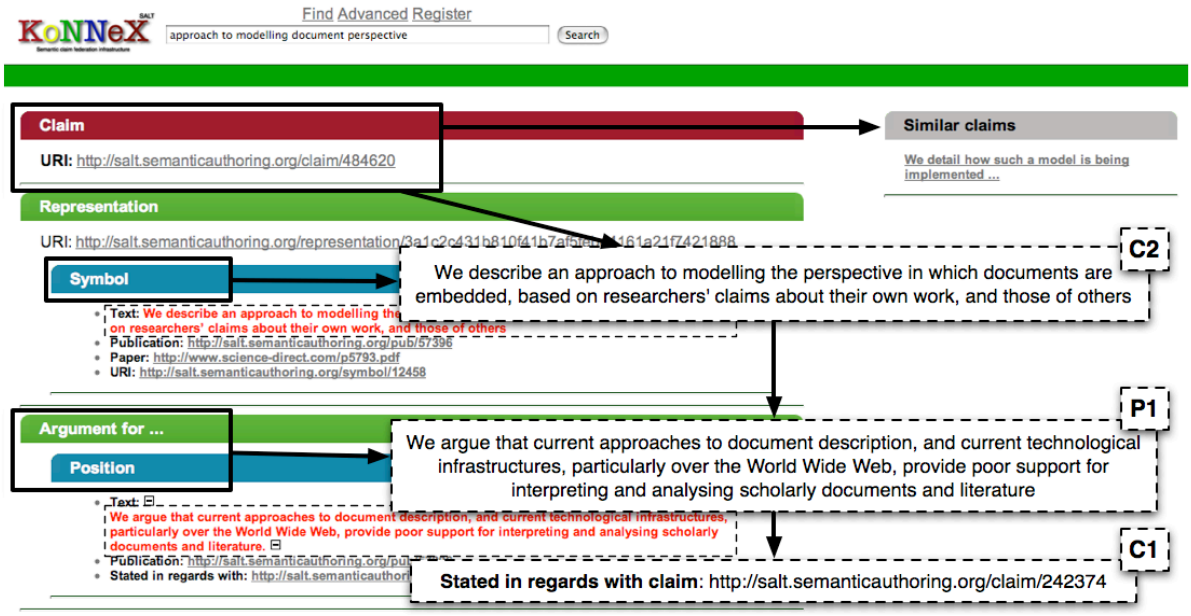



**Figure 13: Browsing argumentative discourse networks in KonneX[SALT]**

# 5 Open research challenges and future trajectories

Despite the multiplicity of the efforts described in this paper, and of others we have not had space to include, a few key themes emerge. It seems essential that, if we want to enable computers to assist scientists with making sense of the vast collections of research papers, we must develop ways to identify key argumentational components, connect them and their supporting evidence across space and time, and present these connected knowledge networks to users. This would form the basis for what Garcia-Castro, *et al.* (2010) call Living Documents. What is still not clear, however, are three things:

1. How such rhetorical markup will be added to scientific discourse. We have shown a number of examples of tools that allow manual or automated markup, but currently none of these is seamlessly integrated with existing authoring or publishing systems and tools. What we need is a group of audacious practitioners who are willing to perform pioneering work in this field, and allow themselves the exertion of dealing with unfinished software and evolving standards to develop a corpus of 'real' scientific work.

2. This is essential to make progress, since the second thing that is needed is information on exactly how these users will use such enhanced data, and which user tasks will be significantly improved or simplified by this annotation. This requires again uptake by a community, willing to work through the problems of treading new territory and opening themselves up to new ways of browsing and digesting scientific literature.

3. Lastly, we need to find resources, and first, the willingness to explore these new avenues on the part of funding agencies, institutions, publishers and libraries. For them, too, the process of handling rhetorically annotated content might be uncomfortably novel at times, and will cost time and effort. New methods of attribution need to be sought, and new ways of validating intellectual property, which rhetorical markup can help enable. Perhaps there will come a time when a researcher can proudly claim that her hypothesis was proven seven times over – or that her research data was used for twenty papers. Such more finely grained attribution could help establish a different way of achieving scientific collaboration, as

authors, readers, and 'users' of each other's statement become, of necessity, more intimately connected.

Incremental steps toward resolving these questions are emerging in the form of new collaborations and workshops. The emerging *Hypotheses/Evidence/Relationships* research community (de Waard, *et al.* 2009; Hyp-ER, 2009) is helping to catalyse research fora, and collaborations such as the recent Cohere/XIP integration (Buckingham Shum, *et al.* 2010). In tandem, the World Wide Web Consortium's *Health Care and Life Sciences Interest Group* now has an ongoing Task to develop usable and sharable models of biomedical discourse on the web, within which several promising new ideas and proposals have been developed.[22] Among these are the AO model for semantic annotation of documents on the Web, previously discussed. Three of the authors of this paper are participants in this Task. A particularly encouraging development has been an explicit recognition by a number of workers on semantic web applications and by textmining researchers that they have important mutual interests and need to collaborate. Another occurrence of importance in the W3C work has been the opening up of work on integrating ontologies of discourse with ontologies of data and computation (see Das et al. 2010 for a slide presentation on the approach). Although the model is currently in draft form it is a promising step toward web-enabled research reproducibility.

# 6 Conclusion

We hope to have shown that there are many strands of work ongoing in the annotation of the rhetoric and argumentation of scientific discourse on the Web. On the one hand, these strands diverge and develop independently, driven by specific project goals and domain applications. On the other hand, a community of practitioners is developing coming at the same issue from different directions: (computational) linguistics, semantic technologies and standards, bioinformatics and medical informatics, and the publishing and digital library communities. Through direct collaborations, connections through standards bodies, and mutual annotation of full-text corpora, the computational and manual annotation approaches are converging. We are excited about the road ahead. With all its challenges, these related lines of research and development offer us a future vision of truly living and evolving, richly-linked and deeply structured scientific documents, with which we fully exploit the inherent power of global social and computational interaction among scientists on the Web.

# 7 References

Aït-Mokhtar, S., Chanod, J-P. and Roux, C. (2002). Robustness beyond shallowness: incremental dependency parsing. Natural Language Engineering, 8(2/3), pp. 121--144.

Benn, N. (2009). Modelling Scholarly Debate: Conceptual Foundations for Knowledge Domain Analysis Technology. Doctoral Dissertation, Knowledge Media Institute, The Open University, UK. Available as: Technical Report KMI-09-04.

Benn, N., Buckingham Shum, S., Domingue, J. and Mancini, C. (2008). Ontological Foundations for Scholarly Debate Mapping Technology. In *Proc. 2nd International Conference on Computational Models of Argument, Toulouse*, ed. P. Besnard, S. Doutre, and A. Hunter, pp. 61–72. IOS Press.

Buckingham Shum, S. (2003). The Roots of Computer Supported Argument Visualization. Chapter 1, *Visualizing Argumentation*, P.A. Kirschner, S. Buckingham Shum, and C. Carr, Editors. 2003, Springer-Verlag: London. p. 3-24.

[22] W3C Health Care and Life Sciences Interest Group: http://esw.w3.org/HCLSIG/SWANSIOC

Buckingham Shum, S. (2008). Cohere: Towards Web 2.0 Argumentation. *2nd International Conference on Computational Models of Argument*, 28-30 May 2008, Toulouse. IOS Press: Amsterdam.

Buckingham Shum, S. and N. Hammond (1994). Argumentation-Based Design Rationale: What Use at What Cost? *Int. J. Human-Computer Studies*, 1994. 40 (4): p. 603-652.

Buckingham Shum, S., Motta E. and Domingue, J. (1999). Representing Scholarly Claims in Internet Digital Libraries: A Knowledge Modelling Approach. *Proceedings of Third European Conference on Research and Advanced Technology for Digital Libraries*, Paris, Sept. 22-24, 1999. (Eds.) Serge Abiteboul and Anne-Marie Vercoustre. Springer-Verlag

Buckingham Shum, S., Motta E. and Domingue, J. (2000). ScholOnto: An Ontology-Based Digital Library Server for Research Documents and Discourse. *International Journal on Digital Libraries*, 3 (3), pp. 237-248

Buckingham Shum, S., Sándor, A., De Liddo, A. and Bachler, M. (2010). Integrating Human & Machine Document Annotation for Sensemaking. *Knowledge Media Institute Seminar*, The Open University, UK (11 Nov. 2010). Replay available at: http://bit.ly/bhtR3u

Buckingham Shum, S., Selvin, A., Sierhuis, M., Conklin, J., Haley, C. and Nuseibeh, B. (2006). Hypermedia Support for Argumentation-Based Rationale: 15 Years on from gIBIS and QOC. In: *Rationale Management in Software Engineering* (Eds.) A.H. Dutoit, R. McCall, I. Mistrik, and B. Paech. Springer-Verlag: Berlin

Buckingham Shum, S., Uren, V., Li, G., Sereno, B. and Mancini, C. (2007). Modelling Naturalistic Argumentation in Research Literatures: Representation and Interaction Design Issues. *International Journal of Intelligent Systems*, (Special Issue on Computational Models of Natural Argument, Eds: C. Reed and F. Grasso, 22, (1), pp.17-47

Burstein, J.C., Marcu, D. and Knight, K. (2003). Finding the WRITE stuff: automatic identification of discourse structure in student essays. IEEE Intelligent Systems, 18(1):32–39.

Ciccarese, P, Ocana M, Chalfont C, Dow E, West A, Clark T. The SWAN Annotation Framework. *Manuscript in preparation*.

Ciccarese P, Ocana M, Das S, Clark T. AO: An Open Annotation Ontology for Science on the Web. Bio-Ontologies 2010, July 9-10 2010, Boston MA.

Conklin, J. and M.L. Begeman (1988). gIBIS: A Hypertext Tool for Exploratory Policy Discussion. *ACM Transactions on Office Information Systems*, 4 (6): p. 303-331.

Corney, D., Buxton, B., Langdon, W. and Jones, D. (2009). BioRAT: extracting biological information from full-length papers. Bioinformatics, 20(17):3206, 2004.

Das, S. 2010. Scientific Discourse: Discourse, Data and Experiment. Presentation available on Slideshare: http://www.slideshare.net/sdas617/sci-discourse-nov-2010

De Liddo, A. and Buckingham Shum, S. (2010). Cohere: A prototype for contested collective intelligence. *CSCW 2010 Workshop: Collective Intelligence In Organizations.* February 6-10, Savannah, GA.

de Waard, A., Buckingham Shum, S., Carusi, A., Park, J., Samwald, M. and Sándor, A. (2009c).

Hypotheses, Evidence and Relationships: The HypER Approach for Representing Scientific Knowledge Claims, *Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, co-located with the 8th International Semantic Web Conference (ISWC-2009).

de Waard, A. Buitelaar, P., & Eigner, T. (2009b), Identifying the Epistemic Value of Discourse Segments in Biology Texts, In: *Proceedings of the Eighth International Conference on Computational Semantics,* Tilburg, The Netherlands, Jan.7-9 2009.

de Waard, A. and Kircz, J.G. (2008). Modeling scientific discourse - shifting perspectives and persistent issues, ELPUB2008. *Open Scholarship: Authority, Community, and Sustainability in the Age of Web 2.0 – Proc. of the 12th Int. Conference on Electronic Publishing*, June 2008, Eds. L. Chan and S. Mornati, pp. 234-245

de Waard, A., and Pandermaat, H. (2010). A Classification of Research Verbs to Facilitate Discourse Segment Identification in Biological Text, *Proceedings of the Interdisciplinary Workshop on Verbs. The Identification and Representation of Verb Features*, Pisa, Italy, November 4-5, 2010.

de Waard, A. and Pandermaat, H. (2009), Categorizing Epistemic Segment Types in Biology Research Articles. *Workshop on Linguistic and Psycholinguistic Approaches to Text Structuring (LPTS 2009)*, September 21-23 2009. – to be published as a chapter in *Linguistic and Psycholinguistic Approaches to Text Structuring*, Laure Sarda, Shirley Carter Thomas & Benjamin Fagard (eds), John Benjamins, (planned for 2010).

de Waard, A. and Tel, G., (2006a). The ABCDE Format: Enabling Semantic Conference Proceedings, In: *Proceedings of the First Workshop on Semantic Wikis, European Semantic Web Conference (ESWC 2006)*, Budva, Montenegro, 2006.

Feltrim, V., Teufel, S., Gracas Nunes, G. and Alusio, S. (2005). Argumentative Zoning applied to Critiquing Novices' Scientific Abstracts. In *Computing Attitude and Affect in Text: Theory and Applications*, James G. Shanahan, Yan Qu, Janyce Wiebe (Eds.) Springer, Dordrecht, The Netherlands, 2005. Pp. 233-245.

Forterre P. (2005). The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. Biochimie. 87(9-10):793-803. Epub 2005 Apr 12.

Garcia-Castro, A., Labarga, A., Garcia, L., Giraldo, O., Montaña, C. and Batemana, J. (2010). Semantic Web and Social Web heading towards Living Documents in the Life Sciences. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8, (2-3), pp. 155-162

Garten, Y., Altman, R. (2009). Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC bioinformatics*, 10(Suppl 2):S6, 2009.

Groza, T., Handschuh, S., Moller, K., Decker, S. (2007a) SALT – Semantically Annotated LaTeX for Scientific Publications. In Proceedings of the 4th European Semantic Web Conference (ESWC 2007), Innsbruck, Austria.

Groza, T., Moller, K., Handschuh, S., Trif, D., Decker, S. (2007b) SALT: Weaving the claim Web. In Proceedings of the 6th International Semantic Web Conference (ISWC 2007), Busan, Korea.

Groza T., Moller,K., Handschuh, S., Decker, S. (2008) KonneX-SALT: First Steps Towards a Semantic Claim Federation Infrastructure. In Proceedings of the 5th European Semantic Web Conference (ESWC 2008), Tenerife, Spain.

Halasz, F.G., Moran, T.P. and Trigg, R.H. (1987). Notecards in a Nutshell. *Proceedings of the SIGCHI/GI conference on Human factors in computing systems and graphics interface*. Toronto, Ontario, Canada: ACM Press: NY. pp. 45–52.

Harmsze, F. (2000). *A modular structure for scientific articles in an electronic environment*. Doctoral Dissertation, University of Amsterdam, NL.

Hilbert, M., Lobin, H., Bärenfänger, M., Lungen, H. and Puskás, Cs. (2006). A Text-technological Approach to Automatic Discourse Analysis of Complex Texts Proceedings of KONVENS 2006

Hoey, M. (1979). Signalling in Discourse. No. 6 in Discourse Analysis Monograph. Birmingham, UK: University of Birmingham.

Hyland, K. (1998). Persuasion and context: The pragmatics of academic metadiscourse. Journal of Pragmatics 30(4): 437–455.

Hyp-ER (2009). *Hypotheses/Evidence/Relationships Workshop*, Elsevier Science, Amsterdam. 11-12 May 2009: http://bit.ly/alkt5e

Hyland, K. (2005). Metadiscourse. Continuum, 2005.

Kircz, J.G.. (1998). Modularity: The next form of scientific information presentation? Journal of Documentation 54: 210–235.

Kuhn, T. S. (1962). The Structure of Scientific Revolutions, Chicago: Univ. of Chicago Pr.

Kunz, W., Rittel, H.W.J. (1970) Issues as elements of information system. Working paper 131, Institute of Urban and Regional Development, University of California.

Liakata, S. Teufel, A. Siddharthan, and C. Batchelor. (2010). Corpora for the conceptualisation and zoning of scientific papers. 2010.

Lisacek, F., Chichestera, C., Kaplan, A., Sandor, A. *Discovering Paradigm Shift Patterns in Biomedical Abstracts: Application to Neurodegenerative Diseases.* In Proceedings of the 1st International Symposium on Semantic Mining in Biomedicine (SMBM), Hinxton, Cambridgeshire, UK, 2005.

Mann, W. C., Thompson, S. (1987) Rhetorical Structure Theory: A theory of text organization. Technical report, Information Science Institute, RS-87-190.

Marcu, D. (2000). The Theory and Practice of Discourse Parsing and Summarization, MIT Press, Cambridge, MA, 2000.

Mercer, N. (2004). Sociocultural discourse analysis: analysing classroom talk as a social mode of thinking. *Journal of Applied Linguistics*, 1(2), 137-168.

Miles, Alistair; Bechhofer, Sean: SKOS Simple Knowledge Organization System Reference, W3C Working Draft 25-January-2008, http://www.w3.org/TR/skos- reference

Raheel Nawaz, Paul Thompson, John McNaught and Sophia Ananiadou (2010). "Meta-Knowledge Annotation of Bio-Events". *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*: Valletta, Malta. pp 2498-2507.

Ogden, C. K., Richards, I. A. (1923) The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism. Magdalene College, University of Cambridge.

Passant A, Ciccarese P, Breslin J and Clark T. SWAN/SIOC: Alignment Between the SWAN and SIOC Ontologies. World Wide Web Consortium, W3C Interest Group Note 20 October 2009. http://www.w3.org/TR/hcls-swansioc/

Pendar, N., Cotos, E. (2008). Automatic identification of discourse moves in scientific article introductions, Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, p.62-70, June 19-19, 2008, Columbus, Ohio

Polanyi, L., Thione, G.L., van den Berg, M. and Culy, Ch. (2004). LiveTree: An Integrated Workbench for Discourse Processing (2004) ACL2004 - Workshop on Discourse Annotation

Rittel, H.W.J., Second Generation Design Methods. *Interview in: Design Methods Group 5th Anniversary Report: DMG Occasional Paper*, 1972. 1: p. 5-10. Reprinted in: *Developments in Design Methodology*, N. Cross (Ed.), 1984, pp. 317-327, J. Wiley & Sons: Chichester

Rittel, H. W. J. and Webber, M. M. (1973). Dilemmas in a General Theory of Planning. *Policy Sciences* 4, 155-169.

Rocher, G and Brown, J. 2009. The Definitive Guide to GRAILS. New York: Apress. ISBN13: 978-1-59059-995-2.

Ruch, P., Boyer, C., Chichester, Ch., Tbahriti, I., Geissbühler, A., Fabry, P., Gobeill, J., Pillet, V., Rebholz-Schuhmann, D., Lovis , C. and Veuthey, A-L. (2007a). Using Argumentation to Extract Key Sentences from Biomedical Abstracts, Int J Med Inform, 76(2-3): 195-200. 2007.

Ruch, P., Geissbühler, A., Gobeill, J., Lisacek, F., Tbahriti, I., Veuthey, A-L. and Aronson, A.R. (2007b). Using Discourse Analysis to Improve Text Categorization in MEDLINE. Medinfo 2007.

Sándor, Á,. Vorndran, A. (2009). An exploratory system for automatic assistance in peer reviewing research articles in educational sciences. NLPIR4DL – Workshop on text and citation analysis for scholarly digital libraries. Workshop at ACL-IJCNLP, 2009. Singapore.

Sándor, Á,. Vorndran, A. (2010). The detection of salient messages from social science research papers and its application in document search. Workshop on Natural Language Processing Tools Applied to Discourse Analysis in Psychology, Buenos Aires, Argentina, May 10-14. 2010.

Shah NH, et al. (2009) Ontology-driven indexing of public datasets for translational bioinformatics. BMC Bioinformatics 10 Suppl 2:S1

Shiffrin, R.M. and Börner, K. (2004). Mapping Knowledge Domains. *Proc. National Academy of Sciences*, 101, pp. 5183-5185. (Special Issue Editorial)

Shotton D (2010) CiTO, the Citation Typing Ontology. Journal of Biomedical Semantics 1 (Suppl 1): S6.

Solov'ev, V. I. (1981). Functional characteristics of the author's abstract of a dissertation and the specifics of writing it. Scientific and Technical Information Processing 3: 80–88. English translation of Nauchno-Tekhnicheskaya Informatsiya, Seriya 1, Number 6, 1981, 20–24.

Swales, J. (1990). Genre Analysis: English in Academic and Research Settings. Cambridge, UK:

Cambridge University Press.

Teufel, S. (1999). Argumentative Zoning: Information Extraction from Scientific Articles. *Doctoral Dissertation*, University of Edinburgh.

Teufel. S. (2005). Argumentative Zoning for improved citation indexing. In ``Computing Attitude and Affect in Text: Theory and Applications'' James G. Shanahan, Yan Qu, Janyce Wiebe (Eds.) Springer, Dordrecht, The Netherlands, 2005. Pp 159-170.

Teufel, S., Siddharthan, A. and Batchelor, C. (2009). Towards Domain-Independent Argumentative Zoning: Evidence from Chemistry and Computational Linguistics. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09), Suntec, Singapore.

| Feature/ Approach | Coarse-grained rhetorical structure | Fine-grained rhetorical structure | Relations | Polarity | Weights | Provenance | Shallow Metadata Support | Domain Knowledge | |
|---|---|---|---|---|---|---|---|---|---|
| **Harmsze** | Modules | Elementary module | Structuring, organisational and discourse | Implicit (within relations) | No | No | Yes | Open | s<br>ir |
| **ScholOnto** | No | Node, Claim | Cognitive Coherent | Explicit (+ / -) | Explicit (1, 2) | Yes (duplicates) | No | Open | G |
| **De Waard** | Rhetorical Blocks | Rhetorical element | Argumentative | Explicit (within the pairs of relations) | No | No | Yes | Open | sc |
| **SWAN** | No | Discourse element, Research statement | Argumentative and Cognitive Coherent | Implicit (within relations) | No | Yes (duplicates) | Yes | Yes (Gene, Protein) | |
| **SALT** | Rhetorical Blocks | Rhetorical element (Nucleus, Satellite, Claim, Support) | Rhetorical and argumentative | Implicit (within relations) | No | Yes (pointers) | Yes | Open | sc |

**Figure 3: Comparative overview of the discourse representation approaches**