Modeling Rhetoric in Scientific Publications

Anita de Waard[1, 2], Leen Breure[1], Joost G. Kircz[3], Herre van Oostendorp[1]
[1] *Center for Content and Knowledge Engineering, Utrecht University, The Netherlands*
[2] *Advanced Technology Group, Elsevier, Amsterdam, The Netherlands*
[3] *KRA Publishing Research, Amsterdam, The Netherlands*

Despite the advent of computer-centered ways of creating and accessing scientific knowledge, the format of the scientific research article has remained basically unchanged. We have developed a model of a more appropriate form for research publications to structure scientific articles, based on a rhetorical structure which is ubiquitous in (natural) science papers. The model has three components: defining rhetorical elements inside the documents, the identification of the argumentational relationships between these elements; and the connection of data elements and entities to external sources.

Keywords: Rhetoric of science, argumentation theory, Story Grammar, science publishing.

1 INTRODUCTION

There is a growing body of work devoted to distilling the knowledge contained within the vast array of scientific literature. In bioinformatics, for example, considerable time and effort is spent in determining the identity of, and relationships between, proteins and genes. Currently, there are two ways to glean these relationships: either by letting human curators read the papers and enter detailed database entries – which has as obvious downside that human work is expensive and doesn't scale. The other route is to let computers extract the names of genes and proteins from biological papers, and combine them to form a database of biological relationships. Of course, what matters is not whether genes and proteins are mentioned in the same paper, or the same sentence, but whether there is an actual relationship between them, that is backed up by research data. We believe there is a third way to determine this relationship: by letting authors explicitly mark up the rhetorical and argumentational structures of their findings and claims during the authoring/editing process, and offering this granular content to a search engine (see e.g. [8]). The purpose of our research is to develop such a model of the scientific article, better suited for usage in an electronic environment.

As has been argued by Latour [10], Kuhn [9] and others, the main purpose of a scientific publication is to convince others of one's viewpoint. If truth – including scientific truth – is decided by the collective, then convincing the collective of the truth of your contribution is the main goal of every author. To optimally represent a scientific paper, we should, therefore, model how it aims to convince. As Latour and Bazerman [1] note, the two main ways in which arguments are usually made in a scientific paper is through references, and by presenting data.

Consider the following sentences, typical for a publication in Cell Biology [15]:

i. **Dhh1p**, an activator of decapping involved in the 5′ to 3′ decay of normal mRNA (Coller et al., 2001),
ii. **accumulated in P-bodies in lsm1Δ, dcp1Δ, and dcp2Δ strains** (Figures 1F, middle panel and left panel and 1G and data not shown).
iii. **This result indicates that the presence of Upf proteins in P-bodies is due to a defect in NMD** and
iv. **implies that NMD targets PTC-containing mRNAs to P-bodies**.

In the first sentence, a statement is made which underlies the rest of the argumentation – namely that *Dhh1p* is '*an activator of decapping....*' The justification for this lies within another publication, *Coller et. al,, 2001*, which the reference tells us is an article in Cell from 2001. To verify statement i, we need to find the Coller paper, read it, and find the place where the role of Dhh1p is discussed. Then, we need to see what other data or further references this statement is built on, by looking at the figures referred to sentence ii. Together, i. and ii. represent the empirical backing for claims iii. and iv[1]. In this way, we reconstruct or 'unfold', to quote Latour, the chain of reasoning preceding this statement. This is what Latour calls the stratification of scientific publications: references are used as shorthand for facts,[2] and links to (images of) data[3] are 'mobilised' to support the author's statement, in 'a folded array of successive defence lines'. We wish to 'unfold' this

[1] With varying degrees of certainty: the central claim in iii is *indicated*, whereas the claim in iv is merely *implied* – see section 2.3 for a further elaboration of such relationships

[2] "The presence of absence of references, quotations and footnotes is so much a sign that a document is serious or not that you can transform a fact into fiction or a fiction into fact just by adding or subtracting references", [10, p. 33]

[3] "Belief in the author's word is replaced by the inspection of 'figures'."[ibid., p. 47]

argumentation stratification, and (allow the author to) create an explicit argumentation structure, where the author explicitly marks which claims s/he makes, and what arguments these claims are based on.

Our goal is to allow the creation of *lines of reasoning* within a text, and between texts, and present the user with a network of linked claims, which lead back to research data. The utopia envisioned is that a reader will be able to assess the validity of a claim, and gain insight into dependencies between claims, and their empirical backing. Using (semantic) web technologies, authoring tools and modeling languages, we hope to build a system that allows the author to create these deconstructed, connected documents. If this system is realized, even in part, it could result in the creation of much more valuable databases of (biological) relationships. Rather than counting how many documents mention two concepts together, we hope to enable an assessment of how many independent measurements actually support a claimed relationship.

2 RHETORICAL MODEL OF A SCIENTIFIC PUBLICATION

2.1 Sketch of the model

Our model has three elements, detailed below:

1. A rhetorical schema, defining the order and rhetorical role of the document sections (section 2.2);
2. An analysis of the argumentation structure of the paper, as reflected in the argumentational relationships between the document sections (section 2.3)
3. The identification of data and entities within the documents (section 2.4).

As an example, we have modelled the backing of sentence i in Figure 1:

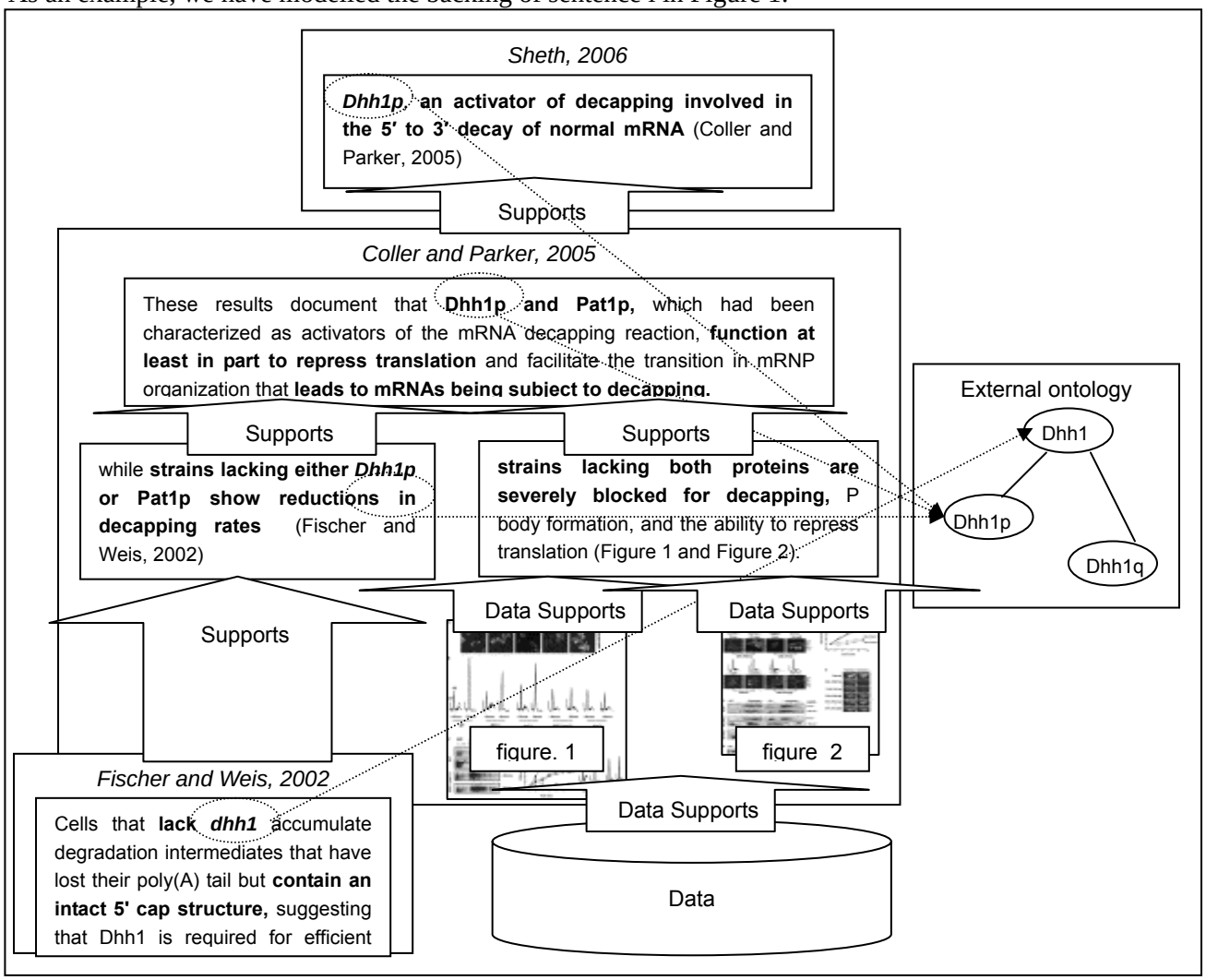

Fig. 1. A model showing parsing of statements between three documents: Sheth (2006), who quotes Coller (2005), and Fisher and Weis (2002), who are quoted by Coller. Arrows represent the argumentative relationships between the document elements (supported by, supported by data), and identification of entities (Dhhp and Dhhp1, connected by dotted lines to the ontology) and data (to the database below).

The specific structure and argumentation of scientific documents changes with the (sub)domain of science. For example, mathematics or statistics can play either an important role in argumentation, or no role at all. The nature of the subdomain can be experimental, theoretical, or computational; the demands of all three are quite different.

As a first approximation to span the vast area of the natural sciences, we have chosen two very different subgenres: cell biology and paleontology. Cell biology is chosen because it is a mature, large field, with very tightly structured publications and very strict, domain-specific criteria for evidence and claim presentation. As a contrast, we will study a corpus of documents in paleontology – both document structure and the type of scientific evidence deemed acceptable here are much more loosely defined, and the field is much more interdisciplinary, situated between (and borrowing scientific conventions from) both geology and archeology.

2.2 Rhetorical Schema

To develop our schema, we compared the sectional divisions proposed by six different approaches to rhetoric modelling, summarized in Table 1:

- Quintillian's 'Institutio Oratori': This classic oratory work proscribes the sections of an oratory, and the roles they play [11].
- Cell style guide: An online manual stating sections that a Cell paper must be subdivided into [3]
- Harmsze: In her thesis, she modelled experimental physics articles to a modular format, where the modules are tailored for multiple use [6].
- Van Oostendorp: A macrostructure of empirical psychology publications was determined [18].
- iPad Schema*:* iPad is an Electronic Lab Notebook–where experiments in biology are noted in XML format, developed by Cognium Systems in France. In particular, we have looked here for the Experiment details for our models[4].
- Story Grammar*:* Story grammars are "schemes for formalising the structure of stories" [12], with the intent of modelling the psychological categories assigned to different parts of a story. They provide a useful model for cognitive functions of story summarisation and comprehension [16].

Quintilian	*Cell Style*	*Harmsze*	*Van Oostendorp*	*iPad Schema*	*Story grammar*		*Our model*	
exordium (Introductio n)	Introduction	Positioning - Situation	Theoretical Background	Project, Participant, funding source	Setting	Time/ Characters/ Location	Introduction	Positioning
narratio (Statement of facts)		Positioning - Central problem	Research question, Hypothesis, Prediction	Central Question, Strategy	Theme	Goal/ Attempt		Central Question, Hypothesis, Strategy
partitio (Outline)	Summary[5]							
confirmatio (Proof)	Experimental procedures	Methods	Method, Input, Organism, Output,	Experimental setup (Goal, Method, Protocol)	Episode 1...n	Subgoal, Attempt,	Experiment 1...n	Setup: Goal, Protocol
	Results	Results, Interpreta-ion	Results, Observations	Experimental results (Data, Interpretation)		Outcome		Results: Data, Interpretatio n
refutatio (Refutation)	Discussion	Outcome	Evaluation		Resolution		Discussion	
peroratio (summary)			Theoretical interpretation, Conclusion				Conclusion	

Table 1. Various rhetorical elements compared to our proposal

The sectional outline is quite comparable between all approaches: specifically, the division into Introduction, (Methods and) Results, and Discussion is omnipresent[6]. We have also proposed this basic tripartite division as a simple RDF-based model[7] for computer science publications, the abcde format [4] which

[4] See http://www.cogniumsystems.com/ for more details on the iPad project

[5] Most instructions require an Abstract, which is usually placed preceding an article, and sometimes replacing it in abstract databases. As a summary/preview at this point in an article, the Article Outline is sometimes used. – see also [1] ch. 4

[6] Also compare sections for this conference: "Abstract/Keywords", "Introduction", "Material and Methods", "Results," "Discussion", "Conclusion"

[7] RDF = Resource Description Framework, the 'lingua franca' of the Semantic Web, see http://www.w3.org/RDF/

is currently being considered as a format for the online workshop proceedings of a Semantic Web conference[8].

The Experiments section forms the 'meat' of a research article; we have used the iPad schema and the concepts of 'Episodes' from Story grammar to come to a division of this section into parts, which contain groups of Goal/Protocol/Data/Interpretation. This is probably where the most fine-grained divisions within articles will come, and also where the most useful results of author-based rhetorical parsing can be gained.

2.3 Argumentation Structure/Coherence relations

Many models of discourse relations have been described in linguistic, philosophical and computer science literature. Hovy [7] has performed an exhaustive study over 35 relation taxonomies and came up with a (statistically motivated) taxonomy of over 120 discourse relations, which he differentiates in three basic kinds: presentational, semantic, and personal. We know of two studies that explicitly concern scientific discourse relations:

- An elaborate link taxonomy devised by Harmsze [6] to classify relationships between modules in physics, based on van Eemeren's pragma-dialectical theory of argumentation [17];
- A discourse ontology developed for the ClaiMaker project [2], an interface to create (post-hoc) argumentation relationships between textual elements in science, partly based on the coherence relations theory by Sanders et. al. [13].

Since our aim is to model realistic scientific reasoning, and, to help the authors, provide as few categories as possible, we have chosen a set of 8 connectors (or four, with their antonyms)[9]:

- proves/refutes;
- supports/contradicts;
- agrees/disagrees;
- suggests/suggests that not.

We especially need to differentiate between 'supports' and 'suggests', since this is a difference that is often encountered in research publications. We will use this as a first-order approximation for the many nuances of certainty which scientists use, such as 'demonstrated', 'examined', 'proposed', 'confirms', 'are suspected to', 'are thought to', 'provides evidence that indicates', 'found', 'convincingly documents', 'surprisingly shows', 'could be a result of', 'are not reflective of', 'importantly, we observed that', 'consistent with', 'strikingly', 'documents that', etc[10]. In using these relations, we furthermore identify whether the links are to other text or texts, or to data, since our goal is to identify relationships leading back to data. We will further examine the validity of this set of relations after modelling a corpus of articles and further exploring it in the light of work on causal connectives in cognitive linguistics, such as that by Sanders et. al. (e.g. [14]).

2.4 Identification of entities and data

The third part element of our proposal to make the research paper computer-processable and enable discovery of and linkage between documents, is to identify the entities and data presented in the document, and relate them to (virtual or real) representations. This means that the author (and/or the system) uniquely identifies any entities which can be classified in a database, such as:

- In Biology: Genes, proteins, cell lines, organs, biological systems, organisms.
- In Palaeontology: Geological time (epoch, period, formation), organism, GPS coordinates, rock type.

Author-supported entity identification should be exact, and enable a data-centered access to the document collection (for example, by allowing a geographical map of digging sites as a way into the paleontology literature). We envisage this to be a computer-supported process; preliminary investigations on the Cell corpus have indicated that entity identification of genes and proteins during the editorial process should be feasible using text-mining techniques to identify Gene Ontology terms [5].

3 DISCUSSION

As a next step, we will model a corpus of articles in the two chosen subdomains, Cell Biology and Palaeontology, and make these available to subject specialists. Critical questions that need to be answered at this stage are:

- Is the basic concept of identifying elements and relating them with argumentation links, useful?
- Are the element sets adequate to cover these two varied types of science publication?

[8] 1st Semantic Authoring and Annotation Workshop, http://saaw2006.semanticweb.org/

[9] Luckily, the wealth of available relation taxonomies provides us with ample relations to choose from, should we need to expand our modest set.

[10] These are just a small subset of the connectors used in the Cell article under scrutiny [15]

– Is the granularity adequate and feasible?
– Is the set of argumentation relations adequate to represent scientific argumentation?

Once the model has been tested (and probably adjusted), the next stage is to find or develop a tool to model this content. We are specifically looking, in this phase, for authoring tools that maintain relations between data sets and identified entities throughout the authoring process, and allow for typed linking between text fragments. Specifically, we will be looking into using electronic lab notebooks as an input system. Increasingly, these e-Lab notebooks use XML as their native format (an example is iPad[5], developed at the Institute Pasteur). We hope to utilise the notebook output to develop an XML schema to code the text fragments and extend it using RDF to indicate the relations between elements, and between text, entities and data.

Critical questions at this stage are:

– Are the tools able to represent the richness of the final model?
– Are authors and editors able, and willing to use it?
– Can we find a visualisation that optimises creation and manipulation of this content?

Once we obtain a suitable corpus, we will test it with users, and investigate an appropriate authoring/editing tool. If our investigations prove worthwhile, Elsevier is interested in using this system in an editorial setting.

ACKNOWLEDGMENT

This work was supported in part by a grant from the Netherlands Organisation for Scientific Research (NWO), under the Casimir programme.

REFERENCES

[1] Bazerman**,** C., Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science. (Madison, WI: Univ. Wisconsin Press, 1988)
[2] Buckingham Shum, Simon J. Uren, V. et. al , Modeling Naturalistic Argumentation in Research literatures: Representation and Interaction Design Issues, Tech Report kmi-04-28, December 2004
[3] Cell, Information for Authors, http://www.cell.com/misc/page?page=authors#SubmissionRA
[4] De Waard, A. and Tel, G., The ABCDE Format, Enabling Semantic Conference Proceedings, Semantic Wiki Workshop, European Semantic Web Conference, 2006
[5] Doms, Andreas and Michael Schroeder. GoPubMed: Exploring PubMed with the GeneOntology. Nucleic Acids Research, 2005 33: W783-W786; doi:10.1093/nar/gki470
[6] Harmsze, F.A.P. PhD Thesis, February 9, 2000. A modular structure for scientific articles in an electronic environment For a PDF version see: http://dare.uva.nl/record/78293 (253 pp).
[7] Hovy, E. Automated discourse generation using discourse structure relations. Art. Intelligence 63(1-2): 1993. 341-386.
[8] Kircz, J.G.. Rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. Journal of Documentation, 47, 4 (December), 1991, 354-372.
[9] Kuhn, Thomas, The Structure of Scientific Revolutions (Chicago: University of Chicago Press, 1962)
[10] Latour, B., Science in Action, How to Follow Scientists and Engineers through Society, (Cambridge, Ma.: Harvard University Press, 1987)
[11] Quintilian, M. F,. Institutio Oratoria , trans. H. E. Butler (London: Heinemann, 1921)
[12] Rumelhart, D. Notes on a schema for stories. In D. Bobrow and A. Collins, editors, Representation and Understanding: Studies in Cognitive Science (New York: Academic Press, 1975)
[13] Sanders, T., W. Spooren and L. Noordman. Toward a Taxonomy of Coherence Relations. Discourse Processes 15: 1-35. (1992)
[14] Sanders, T., and L. Noordman. the role of Coherence Relations and Their Linguistic Markers in text Processing. Discourse Processes, 29(10) 37 – 60 (2000)
[15] Sheth, U. and Parker, R. Targeting of Aberrant mRNAs to Cytoplasmic Processing Bodies, Cell, Volume 125, Issue 6, 13 June 2006, Pages 1095-1109.
[16] Thorndyke**,** P. W. Cognitive structures in comprehension and memory of narrative discourse. Cognitive Psychology (1977), 9 pp. 77–110.
[17] Van Eemeren, F. H,. R. Grootendorst and T.k Kruiger. Handbook of Argumentation Theorie (Amsterdam: Floris Publications, 1987)
[18] Van Oostendorp, H. and Hamaker, C, De Invloed van Explicitering van Tekststructuur op het Onthouden, Tijdschrift voor Onderwijsresearch 3 (1978), nr. 3 {In Dutch}