

The Story of Science: A Syntagmatic/Paradigmatic Analysis of Scientific Text

Anita de Waard
Elsevier Labs, Burlington, VT, USA
Utrecht Institute of Linguistics, Utrecht, The Netherlands
a.dewaard@elsevier.com

ABSTRACT

Following Latour (1987), Latour and Woolgar (1979), Bazerman (1988), and others, we have proposed a model for scientific texts as being ‘stories, that persuade with data’ (de Waard et al., 2006; de Waard et al., 2009; de Waard and PanderMaat, 2009). The persuasive component (how claims are formulated and recognized) is being developed in a number of projects (de Waard et al., 2009); the data component is the object of a great deal of study on various platforms (see overview in de Waard, 2010). In this paper, we wish to comment on the narrative component: how a scientific paper is similar to a fairy tale, and how techniques developed to parse and access fairy tales could be used to improve access to scientific knowledge. This short paper has three parts: first, we provide a brief introduction on fairy-tale structure analysis; next, we offer a small overview of how scientific text can be similarly analyzed, and third, we discuss some ways we could use tools and technologies developed within digital humanities for improving access to scientific knowledge.

1. The Structure of Fairy Tales: Propp vs. Lévi-Strauss

The analysis of narrative structures in folktales has developed in two directions: one, following Propp, is a syntagmatic analysis of story structure, where the chronological order of events unfolding in folktales is described (Propp, 1968). Building on from Propp’s analysis, work in the sixties and seventies by e.g. Thorndyke (1975) and Rumelhart (1977) focused on defining a ‘story grammar’ or ‘story schema’– the ‘systematic assignment of constituent structure’ to stories. The main goal here is ‘to look at a story and to identify the goals, subgoals, the various attempts to achieve the goals, and the various methods that have been employed’ to tell the story (Rumelhart, 1977).

An orthogonal method of analyzing stories follows the work of Lévi-Strauss, and focuses on a so-called paradigmatic analysis of a story, where ‘groups of relations between actors and events are sought throughout the text’ (Lévi-Strauss, 1955). The focus here is to find and group together elements of mythology, also called ‘mythemes’, which occur in different myths and folk-tales.

The two different views can, and have been (Lévi-Strauss, 1955) represented by a two-dimensional coordinate system, where the syntagmatic analysis occurs in (narrative) time, and the other is a ‘paradigmatic grouping’, by grouping together different events related to e.g. marriage, murder, etc.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

Copyright 2010 AMICUS project at <http://amicus.uvt.nl>.

2. The structure of scientific papers: Up and down and side-to-side

2.1 Syntagmatic analysis

We can look at scientific texts in a similar way. To begin with, the Introduction-Method-Results-and-Discussion (IMRaD) structure, or more correctly, the Introduction-Experiments-Discussion structure, closely mimics the basic Story grammar elements of Setting-Episodes-Resolution. If we call the main research question the protagonist of the scientific story, a paper describes its adventures: from being born (through a lack of understanding of current theory) to facing challenges (as different experiments test and explore various characteristics of the research question) to its eventual destiny: usually, in a scientific paper, a part resolution, part transformation of the research question. Like Episodes, Experiments are often mini-stories in themselves, and start with a subgoal, then describe some development, and end with a small resolution, that leads to the next episode. A parallel between Rumelhart’s (1980) story grammar and the structure of a scientific paper is given in Table 1.

Table 1: Comparing story grammar elements with syntagmatic components of scientific text

Setting	Time	Introduction	Background
	Characters		Objects of study
	Location		Setup
Theme	Goal		Research question
	Attempt		Hypothesis
Episode 1	Subgoal	Experiment 1	Subquestion
			Subhypothesis
	Attempt		Method
	Outcome		Data
			Results
			Implications
Episode 2	Subgoal	Experiment 2	Subquestion
			Subhypothesis
	Attempt		Method
	Outcome		Data
			Results
			Implications
Resolution	Outcome	Discussion	Results
	Event		Next steps

Table 2. Discourse analysis of (Louiseau, 2009) showing segmented text, verb form and semantic verb class (for details of the analyses see de Waard and Pander Maat, 2009 and 2010).

Elementary Discourse Unit	Segment type	Verb tense	Verb Class
Though D3 receptor antagonists can enhance cognitive function,	Fact	Present	Cause and effect
their sites of action remain unexplored.	Problem	Present	Cognition
This issue was addressed	Reg-Result	Past Perfect	Discourse verb
employing a model of social recognition in rats,	Method	Gerund	Procedure
and the actions of D3 antagonists were compared to D1 agonists	Method	Past Perfect	Procedure
that likewise possess pro-cognitive properties.	Fact	Present	Properties
Infusion of the highly selective D3 antagonists, S33084 and SB277,011 (0.04-2.5 µg/side), into the frontal cortex (FCX) dose-dependently reversed the deficit in recognition induced by a delay.	Result	Past	Cause and Effect
By contrast, the preferential D2 antagonist, L741,626 (0.63-5.0) had no effect.	Result	Past	Cause and effect
The action of S33084 was regionally specific	Result	Past	Change and Growth
inasmuch as its injection into the nucleus accumbens or striatum was ineffective.	Result	Past	Cause and effect
A similar increase of recognition was obtained upon injection of the D1 agonist, SKF81297 (0.04-0.63), into the FCX	Result	Past Perfect	Procedure
though it was also active (0.63) in the nucleus accumbens.	Result	Past	Cause and effect
These data suggest that	Reg-Implication	Present	Interpretation
D3 receptors modulating social recognition are localized in FCX,	Implication	Present	Properties
and underpin their pertinence as targets for antipsychotic agents.	Implication	Present	Interpretation

2.2 Paradigmatic Analysis

It seems that a syntagmatic analysis of scientific text is quite straightforward. So what would a paradigmatic analysis look like? In essence, Lévi-Strauss performs two actions: the first steps involves 'break[ing] down [a] story into the shortest possible sentences, and writing each such sentence on an index card bearing a number corresponding to the unfolding of the story.' That corresponds to a discourse parsing into smallest constituent units, also called elementary discourse units (edu's) (Marcu, 1999). Although index cards are no longer in fashion, I do apply a small taxonomy of segment types to elemental discourse units, and the segment types consist of functions of with specific subjects. See table 2 for an example of a fragment of this analysis.

Lévi-Strauss notes that 'a certain function is predicated to a given subject'; 'Or, to put it otherwise, each gross constituent unit will consist in a relation.' Lévi-Strauss' main claim is now that 'the true constituent units of a myth are not the isolated relations but bundles of such relations.' Is this also the case of scientific text?

In scientific discourse, the predicates studied are first of all verbs: their form (tense, aspect, mood etc.) (de Waard and Pander Maat, 2009), what semantic class they belong to (de Waard and Pander

Maat, 2010), and secondly modality markers (e.g., hedging markers (possibly, certainly), modal auxiliaries (might, could) and 'suggests that'-type of constructions). Indeed, the results are similar to Lévi-Strauss' conclusions: we can group these segment types into broader categories. Specifically, it seems that there are three types of text in scientific papers: one type pertaining to conceptual claims and statements; one type pertaining to experimental methods and findings, and one type that contains the 'connecting text', intra- and intertextual elements of the type 'as we have shown' or 'see figure 2', or 'as Lendvai (2008) has indicated', on the one hand, and regulatory or connective segments such as 'these results suggest that', 'from this, it can be deduced' etc.

An attempt at a paradigmatic representation is given in Table 3, where we differentiate between conceptual and experimental discourse. Support for this split is given by the use of verb tense: there is a good correlation with past tense for experimental discourse, and present (modal or direct) forms for both conceptual discourse, and simple present tense for connecting discourse (de Waard and Pander Maat, 2009). A similar overlap with verb form exists (de Waard and Pander Maat, 2010).

Table 3. The segments from Table 2 ordered ‘paradigmatically’ by topic: concept/connection/experiment.

Conceptual discourse	Connecting discourse	Experimental discourse
Though D3 receptor antagonists can enhance cognitive function,		
their sites of action remain unexplored.		
	This issue was addressed	
		employing a model of social recognition in rats,
		and the actions of D3 antagonists were compared to D1 agonists
that likewise possess pro-cognitive properties.		
		Infusion of the highly selective D3 antagonists, S33084 and SB277,011 (0.04-2.5 µg/side), into the frontal cortex (FCX) dose-dependently reversed the deficit in recognition induced by a delay.
		By contrast, the preferential D2 antagonist, L741,626 (0.63-5.0) had no effect.
		The action of S33084 was regionally specific
		inasmuch as its injection into the nucleus accumbens or striatum was ineffective.
		A similar increase of recognition was obtained upon injection of the D1 agonist, SKF81297 (0.04-0.63), into the FCX
		though it was also active (0.63) in the nucleus accumbens.
	These data suggest that	
D3 receptors modulating social recognition are localized in FCX,		
and underpin their pertinence as targets for antipsychotic agents.		

2.3 Combination/contrast

If we combine these two analyses, we can paint a two-dimensional picture of scientific discourse, where syntagmatic and paradigmatic elements are depicted on orthogonal axes. For scientific text, the same dichotomous analyses can be done: the narrative form of subsequent sections of a paper has a parallel to story grammars or schema's, and the topics can be mapped to paradigmatic categories, as shown above. In Fig. 1, the topical vs. sequential axes represent the story grammar-like components versus the three realms discussed in 2.2.

However, there is a second difference between the Proppian and Lévi-Straussian analyses: Propp looks at supersentential units of text (as do we in Figure 1), whereas Lévi-Strauss calls us ‘break down [a] story into the shortest possible sentences’ – that is, to divide it into clauses; which is what I attempt in my segment-type analysis. Figure 2 is a sketch of such a clause-level analysis in terms of discourse flow (x-axis) and in terms of subject type (y-axis), for the text in Table 3. This text is taken from the Introduction, and it is clear the coarser-grained annotation from Figure 1 is an inadequate representation of topics covered in the text: the finer grain is needed (see also Nawaz et al, 2010).

3. How can narrative analysis help scientific understanding?

From the previous, it seems that we can analyze scientific text using insights obtained from narrative analysis. How can this analysis can help improve access to scientific knowledge? Despite the obvious parallels between stories and scientific papers, there is no ‘story grammar’ defined for scientific papers, as such. Various publishers have different schema's for defining papers: see Table 4 for an overview of 4 publishers' subject headings. There have been past efforts to propose a modular structure for scientific papers (Kircz and Harmsze, 2000), and a proposal for a LaTeX-based authoring tool for simple narrative scientific text, the abcdx format (de Waard and Tel, 2006), but currently, there is no consensus between publishers to obtain a finer-grained, rhetorical structure markup, despite some efforts in this direction (e.g., Groza et al , 2009). Hopefully, utilizing similar steps as followed in defining the Proppian markup standards (Malec, 2004) and corresponding markup (Lendvai et al, 2010 a&b), data standards (Declerck et al., 2010) and ontologies (Peinado et al. 2004) could help establish a robust format for scientific narrative markup. Similarly, experiments in

automated markup of these structured fairytales such as AutoPropp (Malec, 2010) as well as other work in automate story generation (e.g. Callaway and Lester, 2002, and references therein), might help identify core elements in scientific text. It would be very interesting to combine tools focusing on this syntagmatic analysis with the perpendicular task of finding common patterns within scientific text; work on e.g. BioEvent extraction (Naw'az et al, 2010) points in this direction, and collaborations on this front at a more granular level might offer a promising way forward.

Table 4. Article sections and DTDs for several publishers.

Publisher/DTD URL	Sections
Biomed Central http://www.biomedcentral.com/xml/dtdtaggingspec.html	Background Results Discussion Conclusions Methods
Elsevier http://www.elsevier.com/frame/work_authors/DTDs/ja50_tagbvtag5.pdf	Introduction Results Discussion Experimental Procedure
Nature Publishing Group ?	Background Findings Discussion Methods
National Library of Medicine http://dtd.nlm.nih.gov/articleauthoring/tag-library/	Intro Results Discussion Conclusions Methods Cases Materials Subjects Supplementary-material
Society for Neuroscience ?	Introduction Result Discussion Materials & Methods

4. References

[1] Bazerman, C. 1988. Shaping written knowledge: the genre and activity of the experimental article in science, Madison, Wisconsin: Univ. of Wisconsin Press, 1988.

[2] Callaway, Charles B., James C. Lester, Narrative prose generation, Artificial Intelligence, Volume 139, Issue 2, August 2002, Pages 213-252.

[3] de Waard, A., and Pandermaat, H. (2010b). A Classification of Research Verbs to Facilitate Discourse Segment Identification in Biological Text, Interdisciplinary

Workshop on Verbs. The Identification and Representation of Verb Features, Pisa, Italy, November 4-5 2010.

[4] de Waard, A. (2010). From Proteins to Fairytales: Directions in Semantic Publishing. IEEE Intelligent Systems 25(2): 83-88 (2010)

[5] de Waard, A. and Pandermaat, H. (2010a), 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).

[6] de Waard, A., Simon Buckingham Shum, Annamaria Carusi, Jack Park, Matthias Samwald and Ágnes Sándor. (2009). 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).

[7] de Waard, A. (2007). A Pragmatic Structure for the Research Article, in: Proceedings ICPW'07: 2nd International Conference on the Pragmatic Web, 22-23 Oct. 2007, Tilburg: NL. (Eds.) Buckingham Shum, S., Lind, M. and Weigand, H. Published in: ACM Digital Library & Open University ePrint 9275.

[8] de Waard, A. and Tel, G., (2006). 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.

[9] de Waard, A. Breure, L. Kircz, J.G. Oostendorp, H. van (2006). Modeling Rhetoric in Scientific Publications.

[10] Current Res. in Inf. Sci. and Techn. pp. 352-356, 2006.

[11] Declerck, Thierry Kerstin Eckart, Piroška Lendvai, Laurent Romary, Thomas Zastrow (2010). Towards a Standardized Linguistic Annotation of Fairy Tales, in: LREC 2010, Language Resource and Language Technology Standards – state of the art, emerging needs, and future developments.

[12] Groza, T., Handschuh, S. Clark, T., Buckingham Shum, S. and De Waard, A. (2009). A Short Survey of Discourse Representation Models, ISWC Workshop on Scientific Discourse Representation, 2009.

[13] Kircz J.G. and F.A.P. Harmsze (2000), Modular scenarios in the electronic age. Conferentie Informatiewetenschap 2000. Doelen, Rotterdam 5 april 2000. In: P. van der Vet en P. de Bra (eds.) CS-Report 00-20. Proceedings Conferentie Informatiewetenschap 2000. De Doelen Utrecht (sic), 5 april 2000. pp. 31-43.

[14] Latour, B. (1987). Science in Action, How to Follow Scientists and Engineers through Society, (Cambridge, Ma.: Harvard University Press, 1987)

- [15] Latour, B., and Woolgar, S. (1979). *Laboratory Life: The Social Construction of Scientific Facts*. Beverly Hills: Sage, 1979.
- [16] Lendvai, P., T. Declerck, S. Darányi, S. Malec (2010a). Propp revisited: integration of linguistic markup into structured content descriptors of tales. In *Digital Humanities 2010*. London, United Kingdom, Oxford University Press, July 2010.
- [17] Lendvai, P., T. Declerck, S. Darányi, P. Gervás, R. Hervás, S. Malec, and F. Peinado (2010b). Integration of linguistic markup into semantic models of folk narratives: The fairy tale use case. In *Proceedings of LREC, 2010*.
- [18] Lévi-Strauss, Claude. (1955). The Structural Study of Myth. *The Journal of American Folklore*, Vol. 68, No. 270, Myth: A Symposium (Oct. - Dec., 1955), pp. 428-444
- [19] Loiseau, F., Millan, M.J. (2009). Blockade Of Dopa-mine D3 Receptors In Frontal Cortex, But Not In Sub-Cortical Structures, Enhances Social Recognition In Rats. *European Neuropsychopharmacology - January 2009 (Vol. 19, Issue 1, Pages 23-33)*.
- [20] Malec, S. A. (2001). Proppian structural analysis and XML modeling. In *Proceedings of CLiP, Duisburg, Germany, December 6-9, 2001*.
- [21] Malec, S.A. (2010). AutoPropp: Toward the Automatic Markup, Classification, and Annotation of Russian Magic Tales, This Workshop (*Amicus Workshop '10*, October 21, 2010, Vienna, Austria)
- [22] Marcu, D. (1999) A decision-based approach to rhetorical parsing, *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, p.365-372, June 20-26, 1999, College Park, Maryland.
- [23] Nawaz, R., Thompson, P., McNaught, J. Ananiadou, S. (2010). Meta-Knowledge Annotation of Bio-Events. In *Proceedings of LREC 2010*, pages 2498-2505.
- [24] Peinado, Federico, Pablo Gervás, Bel'en D'iaz-Agudo, (2004). A Description Logic Ontology for Fairy Tale Generation, In *Fourth Int. Conf. on Language Resources and Evaluation: Workshop on Language Resources for Linguistic Creativity, 2004*
- [25] Propp, V. J. (1968). *Morphology of the folktale*. University of Texas Press: Austin, 1968. (Transl. L. Scott and L. A. Wagner).
- [26] Rumelhart, D. 1975. 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)
- [27] Thorndyke, P. W. 1977. Cognitive Structures in Comprehension and Memory of Narrative Discourse, *Cognitive Psychology* 9, 77-110 (1977)

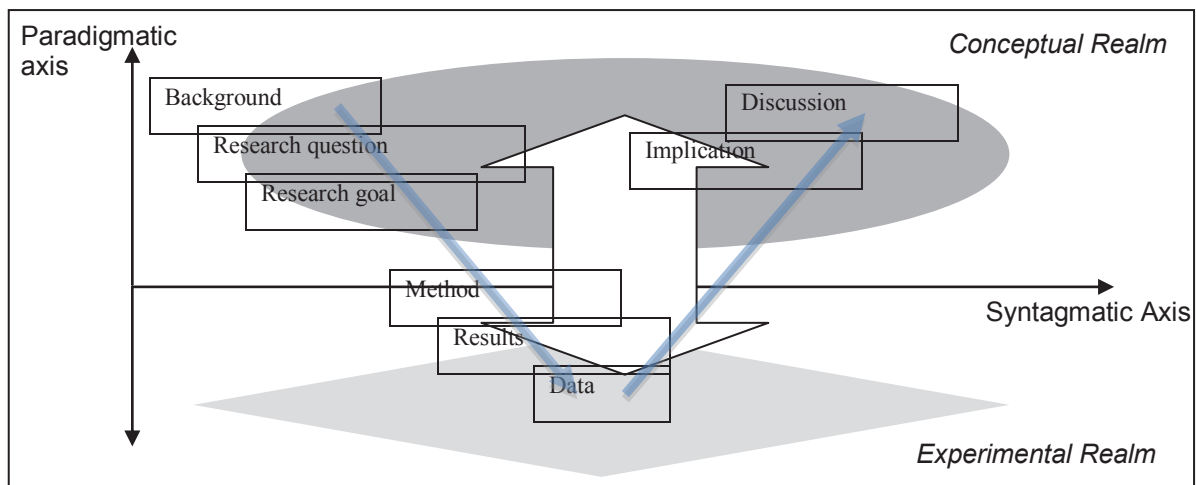


Figure 1. Two axes of analysis: on the x-axis, contiguous structural elements of a research paper; on the y-axis, the main conceptual categories that experimental research covers.

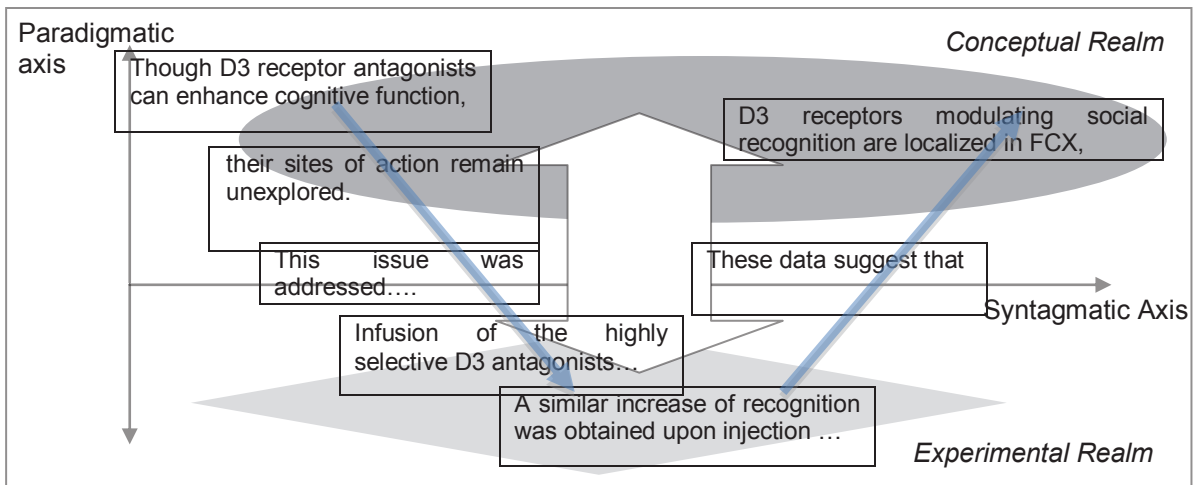


Figure 2. Similar axes as in Figure 1. Note finer-grained textual elements, showing that the move from concept to experiment and back to concept is a recursive pattern that occur at different granularities of discourse.