# RHETORICAL STRUCTURE OF SCIENTIFIC ARTICLES:THE CASE FOR ARGUMENTATIONAL ANALYSIS IN INFORMATION RETRIEVAL

JOOST G. KIRCZ

*Elsevier Science Publishers. P0 Box 103,1000 AC Amsterdam*

In this paper the extent to which modern indexing and information retrieval research meets the needs and requirements of different types of readers is criticised. A review of the stagnation in this field gives evidence for the need for a radically different approach. The main problem is identified as the assumption that knowledge contained in a scientific article can be represented by a semantic network only, and therefore can be manipulated by formal logic approaches. Complementary to this, a plea is made to start an argumentational analysis of the - highly structured - corpus of scientific articles (mainly in physics). Such an analysis might lead to an argumentational syntax which will also enable the non-expert to browse through large quantities of electronically stored articles. A first attempt at such an approach is given. Furthermore the possible use of the Standard General Markup Language (SGML) approach in relation to a hypertext environment for a possible application is discussed.

## 1. INTRODUCTION

THIS ARTICLE DEALS with the problems of automated bibliographic and full text storage of scientific articles. A critical review is given of the stagnation in the field of information retrieval (IR), and a new approach, complementary to existing semantic methods, is suggested.

With the ever increasing number of scientific papers published per year, automated indexing, storage and retrieval systems are indispensable. The traditional methods of indexing and classifying articles used at present in these automated systems are inadequate because of the enormous amount of material available. Any overview with respect to the content of articles gets lost. Different approaches are needed to assist the scientist in finding his or her way through the jungle of published material. Generally speaking, two types of information identification are normally used: (a) bibliographic information: Who (wrote the article), Where (was the research performed), Wherein (were the results published), and When (was the paper published); and (b) scientific classification: What (was calculated or measured), Why (was this subject of interest) and Which (methods were used)? Straightforward bibliographic information is the easiest part to handle in a search attempt for scientific knowledge. Indexing, with regard to content, however, digs deep into the problems of meaning and understanding.



Journal of Documentation. vol. 47, no.4, December 1991, pp. 354—372

All large scientific information handling systems are based, not on the manipulation of the articles themselves, but exclusively on their bibliographic and subject identifiers. This approach limits enormously the disclosure possibilities of written communications. Only well-defined 'search strategies' are feasible and browsing is completely impossible. As a result the current automated systems serve only those people with well defined queries. The interesting point however is, how do we and people who are not (yet) experts in the present (sub)discipline exploit the full knowledge presented in a scientific article? To this end totally new approaches have to be developed which have to be much broader than the reductionist practice of representing the contents of the works solely in bare identifiers (for example lexical notions) and which will also have to contain the argumentational aspects of the scientific discourse.

In this article the non-specialist, browsing, uninformed scientist as reader is taken as the starting point, the person who simply wants to come across what he or she does not know yet. Formalised index terms do not convey much then, so the reasoning of why and how something has been done becomes more important. In order to meet these needs it is proposed that a new net has to be spanned over the document, this time not a set of semantic equivalents, but an argumentational or rhetorical network.

In order to keep the investigation of the problems within certain limits, the familiarity of the present author with the field of physics is taken as a starting point, in the sense that only scientific texts in the natural sciences, particularly physics, and written in English are taken into account. The difference in structure of articles in sciences other than physics, and certainly the difference in writing in other languages (see Clyne [1] for example for the difference between German and English texts) can be very large, and will not be treated here.

After dissection of the different roles a scientific article plays for authors (section 2.1) as well as for the various types of readers (section 2.2) and the form of the article as a function of its role (section 3), the actual state of the art of indexing (section 4) and a critical overview of IR technology (section *5)* are given. It is proposed (sections 6 and 7) that next to the attempts to 'map' knowledge in static frames of well defined notions, the dynamics of the argumentation especially can be used as a most useful tool in disclosing texts. At the end of the paper a possible technical solution for an argumentational network approach is suggested, in order to emphasise the feasibility of such an approach.

The ideas set out in this paper should be seen as a call for the discussion necessary to develop these novel lines further.

## 2. THE ROLE OF THE SCIENTIFIC ARTICLE

The research article in physics plays a crucial role for scientists. Scientists use articles in two completely different ways. On the one hand they are authors, which imposes on them a series of very strict demands in writing up their research and, on the other hand, they are readers of other people's material.

and, in that capacity, are generally not interested in everything that is written. As readers they will sometimes use a very long article only to pick out one single number or half of an idea.

*2.1 The article from the author's point of view*

For the author the article plays different roles. In the first place, the author informs the scientific community of the results of a particular piece of research. The author reports his or her research in a narrative sense in such a way that a non-expert, but trained, colleague will understand the work and will, in principle, be able to repeat the experiment, the calculations or the theoretical work. For those purposes the style of the paper must have explanatory or even tutorial aspects. Secondly, the article is also evidence of the scientific quality and ability of the author and will (re)establish his scientific prestige in the eyes of peers, colleagues and sponsors. Therefore style elements and comprehensiveness must have convincing power.

*2.2 The article from the reader's point of view*

From the reader's point of view the scientific article serves very different goals depending on his or her position and interests. Generally speaking, the article can be defined as 'a source of information which might contain something useful for me', where this 'something useful' can range from the correct address of a co-author to the understanding of an elaborate theory. The article can be used for the retrieval of factual information, as well as providing a browsing researcher with a new idea. Contrary to what one would expect, in analyses of primary scientific publications, a distinction between the various forms of readership is rarely made. However, it is crucial for a good understanding of the role and function of the scientific paper to look at the needs and behaviour of different kinds of reader who can be classified in four broadly defined groups.

*2.2.1 The non-reader*

Interested non-readers are an important group consisting mainly of administrators only concerned by the fact that a particular person or research group has published, where, how often, and what type of publication (for example review, book, research note).

For this group of users of scientific articles, the usual bibliographic information (author's name, journal name, volume, issue and page numbers, etc.) is normally sufficient. The article serves a role in essentially bookkeeping operations and in most cases does not require reading at all.

*2.2.2 The informed reader*

The informed reader is a scientist in the same field as the author, or has at least a good understanding of the field. This type of reader knows what he or she is looking for and is able to find his or her way in the literature quickly. The authors' names are known. Keywords, which are not always very precise, are complemented with background knowledge. This type of reader selects a

limited number of journals and scans the contents or flips through the journal issues. A study of how this type of goal oriented scientist reads journal articles has been performed by Bazerman [2]. The queries are clear and in online bibliographic searches, the standard techniques are sufficient most of the time. The narrow selection of articles retrieved by the informed reader just serves the direct, known, information needs. Only the hard core of information contained in the article is used and there are hardly any reasons to read the article in its entirety.

*2.2.3 The partially informed reader*

The partially informed reader is someone who is starting up in a field, or is just generally interested in adjacent fields to his or her own. Names are sometimes known from review articles or references, but their importance is not fully appreciated. Selections on keywords normally give rise to too many hits in online searches. It is very difficult to find the needle in the haystack. The only hope is to find a good recent review article and to trace back the references. Using a citation index database one is also able to search forward by retrieving those articles which refer to a particular 'old' article[1] . A recent relevant study for the social sciences is given by Ellis [3, 4]. The articles are partially read with emphasis on the general approach, their relation to other (known) work and the conclusions.

*2.2.4 The uninformed reader*

In the last category we find the most interesting group of readers for the present study: the uninformed readers, those who want to learn something new. Via colleagues or by reading news stories in the press, interest in a new field is being aroused. Questions such as, 'What are they doing in high-temperature superconductivity?' or 'What type of methods exist in data reduction?' and so on, are typical for this group. The question is not precise. Any question in return (by living or silicon chip based expert systems) which tries to explicate or interpret the original question (for example 'Do you mean, binary data reduction?'), may be equally answered with yes or no. What is being sought is an understanding, not a number or a firm, formal answer. The articles found are mostly read for their introductions, their overview figures or graphs, their conclusions and their lists of references.

It Is important to note that every scientist on different occasions will play all these four roles of readership depending on the stage of the research or teaching programme. At the same time a particular article can serve all four groups of readers, although not necessarily equally well. Obviously the retrieval requirements for the four categories are very different.

[1] This last way of searching is not a common practice due to costs and the lack of integration of the main bibliographic databases with citation databases.

Although the scientific article has very distinct roles for author and reader as described above, the content is almost completely determined by the author. The reasoning of the article is defined by the author's point of view at the time of writing. It is the author who wants to get his or her message across, which may be a claim, a proof that research is going on or a complete report. The structure of the research article has been developed according to this need (for historical information see also Mullins *et al.* [5]. The author wants to convince the informed reader, who is after all the only one who is able to assess the validity of the work.

Perelman and Olbrechts-Tyteca [6], when discussing in depth the argumentational aspects of a scientific discourse, clearly stress the historical aspects and the role in which the approach by formal logic becomes hegemonic. The line of argumentation in a scientific article is nowadays standardised to a high degree - hence the availability of different types of style manuals and writer aids. An article has to (dis)prove or comment on an idea or experiment following a well defined path. It therefore becomes perfectly acceptable that, for example, Day [7, p. v] can state in his book *How to write and publish a scientific article:* 'Yes, this *is* a cookbook'.

The research article in physics conforms to a strict pattern giving uniformity to every work published and therefore the opportunity to judge the work on the basis of more or less standardised criteria. The work is written for a well defined public (namely the informed reader) and fits in a continuing stream of similarly conceived publications. 'It is true that these authors when addressing a learned society, or publishing an article in a specialised journal, can afford to neglect the means of entering into contact with their public, for the indispensable link between speaker and audience is provided by a scientific institution, the society, or the journal. In such cases, the author has merely to maintain, between himself and the public, the contact already established by the scientific institution' [6, p. 18]. Therefore, adoption of this fixed structure can generate beautiful fake articles which are sometimes jokingly published [8].

This standardised structure also allows sociologists and linguists to analyse or to describe the scientific article; for physics see, for example, Pinch [9] and Bazerman [10]. In all these studies, it is the individual author who is central and whose utterances, projected onto paper, using established norms, are being interpreted. The actual form given is one in which authors, analysts and readers have to confront themselves. The 'why' of the established structure is hardly ever discussed. Nevertheless it is the line of argumentation which determines the convincing power of the work. It is the established facts or theories, the accepted or to be proven approximations and the (hidden) presumptions which, neatly juxtaposed, give the paper a natural and logical appearance. This very straightjacket however enables us to do more with the written text than is currently the practice. Construing the argumentational lines of reasoning might reveal the handles through which the uninformed reader might obtain an understanding of a new field of interest (see section 6).

## 4. THE INDEXING OF THE SCIENTIFIC ARTICLE

Owing to the established form in presentation and the use of discipline dependent jargon, codification in indexing and retrieval techniques can also be established.

In most cases, one tries to catch the themes by assigning terms which represent the contents in a simplified (or, if you wish, generalised) way. Complicated expositions are reduced to a limited number of concepts, or in other words, the content of the article is projected or mapped onto a fixed set of manipulable semantic entities.

Three types of identifiers can be distinguished:

1. The information which in the standard form is transmitted with the paper itself, like the author's name, and all further bibliographic information. This type of inherent, directly available information we will call 'internal' identifiers. In this paper, we will not elaborate much further on it.

2. In order to reduce the total amount of information comprising the full article as well as to link concepts used to related, unused concepts, indexes of keywords (and, for example, chemical registration numbers) are added after the article is published. Such indexes are brought in from outside and thus it is reasonable to call them 'external' identifiers. In section 4.1 they will be discussed further.

3. Although not yet commonly used for retrieval purposes, the bibliographic references at the end of a paper serve as a third identifier. These are also known as citations. As these identifiers do refer to other work related to the actual paper, we will call them 'transmittal' identifiers. In section 4.2 more attention is given.

In normal practice papers are represented by certified lists of internal and external identifiers. The reader can search for relevant material by using (and combining) these lists. Most of the time the lists consist of single worth or small noun phrases. A novel option of a totally different kind is the representation of the paper in so-ca11ed Standard Generalized Markup Language (SGML), wing syntactical entities. Applying SGML one obtains a structured way of indexing, which for that reason is of great value for the present study. SGML entities maybe identical with some identifiers mentioned above, but can also represent whole parts of a text. A full discussion is given in section 4.3.

### *4.1 External identifiers*

Although the article can be considered as being the author's best account of the research performed, the language used by the author, for example in the



Journal of Documentation. vol. 47, no.4, December 1991, pp. 354—372

title or even the abstract, can be totally unclear or confusing.[2] The existence of professional indexers and abstractors proves that even for the (partially) informed reader the paper might be not complete enough. The task of these professionals is to enhance the accessibility of the author's work by adding a limited number of (certified) external identifiers. That way they link the information given to an existing body of already structured identifiers (for example, if an article is on 'superfluid Helium 3', the indexer might add terms like 'phase transitions' or 'cryogenic techniques' etc.). In that sense the indexing adheres firmly to the message (and its direct surroundings) which the author wants to disseminate and not to what a reader might pick out of it, independent of the intentions of the author.

A key difference in indexing for automated systems (online databases) in contrast to classic systems (index card files) is that in classic indexing the index is prepared in final form, or pre-coordinated. While using an automated system the user can join different index terms freely by using all kinds of retrieval software, and the index terms are post-coordinated. In automated systems, by virtue of the capacity of the searcher to manipulate terms, the entire responsibility for retrieval is shifted to the reader. Although one is easily able to browse through journals, browsing through the alphabetic lists of words of electronic indexes is only possible if one knows the field very well.

*4.1.1 Computer assisted indexing*

As the indexing problems increase, given the enormous growth of the number of articles published, attempts are made to speed things up by automatic indexing. Extensive discussions are presented by- van Rijsbergen [11, chapter 2] and Salton and McGill [12, chapter 31. Only a few working systems, to assist the human indexer, are implemented. A good example is the physics database of the Fachinformationszentum-4 Karlsruhe, Germany (for the latest description see Biebricher *et* at [13]).

Automatic indexing is a tool for generating indexes which relies on a series of important assumptions, which inherently limit their usefulness. First of all, it assumes that all relevant information (although not necessarily in its final form) is contained in the text as the author provides it. Second, it assumes that a word used by the author is a measure for the information contained in the work. Third, it is assumed that appropriate weighting functions for various words can be defined in order to rate the different words according to relevance. Fourth, it is assumed that the reader is able to link various index terms with the use of query languages into concrete questions in order to obtain a listing of relevant documents out of the database. It is here where we most strongly see the attempt to squeeze IR into simple data retrieval, a

[2]A good illustration is an article by P. Ginsparg in the top journal *Nuclear Physics B. B295* (FS2I), 1988, 153—170, where the full title reads 'Curiosities at c= 1'. Besides the intriguing vagueness of this title, an extra problem is introduced as the parameter *c* is not the velocity of light. which is often taken as being equal to 1 in this field of research, but the charge on a body in a particular theoretical model.

practice Blair [14] takes as the starting point of his criticism of modem IR techniques.

It is already clear from the first inspection of the presuppositions of automatic indexing that some very important implicit assumptions also hide behind this method, the most important of which is, of course, that we are dealing with a stable situation. The author knows what to report and the reader knows what to look for. The field is also considered mature in the sense that a common language is used throughout the corpus of stored documents in the database and the goal of the research is universally accepted.

As the words are extracted from the work, use is restricted to informed readers, thus making it very difficult for partially informed or uninformed researchers in that field to scan the literature in search of some useful concepts or techniques.

On top of that, in all the writing on automatic indexing hardly any attention is paid to the fact that although English is now the *lingua franca* of science, this does not mean that the word choice is equal for all authors and readers. Automatic indexing is of course never able to spot metaphorical use of words (even if the index is linked to dictionaries) nor, and even worse, anaphora (see Liddy *et al.* [15,16] for research on anaphora in the abstracts of scientific articles). It will also be clear that co-word analysis, where one does investigate the relative distance between two words in order to rate their values better cannot escape from these fundamental weaknesses.

Finally the assumption that the reader/searcher will be able to mix the correct terms into a relevant query also presupposes that the reader and author are more or less within the same group and/or on the same level of information and understanding. This all just proves that as long as one knows where to go, one cannot get lost

*4.2 Transmittal identifiers*

Next to the internal and external identifiers, the corpus of references at the end of an article are also useful identifiers. As they refer not to the article itself but to another article used, or necessary to understand the present one, we call them transmittal identifiers as they transmit rather than assign information.

As it is generally understood by many people (but far from always true) that the more often an article is referred to in scientific literature, the more important the work is, the use of references becomes increasingly political. A good overview of the weaknesses of the use of citation analysis is given by MacRoberts and MacRoberts [17].

*4.3 SGML coding*

A representational approach which emerged not from the library world but from the graphics and computer industry, is the Standard Generalized Markup Language (SGML) method of describing documents.

In SGML two historical approaches come together on the one hand developments in the wordprocessing industry towards a descriptive markup, that is to say a structured system of tags for the identification of typographical

layout: on the other hand developments towards a generic markup in the printing industry, where methods emerged to identify and manipulate major components of texts (headings, lists, highlighted textparts etc.) in order to use them at various times in different documents or in different document presentations (for example to keep the integrity of standard clauses in a great variety of manuals).

SGML is a meta language that describes the syntax of the identifiable entities in the text (see Baron [18] for full treatment).

The unique point is that the approach is, in principle, content free. No 'meaning atoms' are indexed but rather the essential, structural ingredients which comprise the text. The SGML standard description of text for storage in electronic media is independent of the hardware or software being used for data acquisition, as well as independent of the form in which it will ultimately be presented (for example printed or displayed on a video screen).

SGML has become more widespread and popular, and in 1988 international standards were adopted (ISO 8879 [19] and ISO/IEC TR 9573 [20]). Recently a most readable and extensive author's guide to SGML was published by Bryan [21]. Kircz and Bleeker [22] developed a model for the use of a relational database based on SGML, for scientific publishing.[3] The great value of this type of approach to indexing or 'mapping' text is that it does not try to interpret or mirror the thoughts or experimental results of an author as is the case with all types of external identifiers. This method enables the identification of categories which intrinsically makes it possible to retrieve only the section 'Conclusions' of a collection of articles, even if this section is called 'Theoretical implications', 'Final remarks' or whatever. Instead of trying to represent the knowledge contained in a text, in another — shorthand — representation, an attempt is made to dissect the narration into elementary recognisable entities.

## 5. RETRIEVAL TECHNIQUES

In this section a short critical overview of the limitations in retrieval capacity as a result of the indexing methodology described above is given, in order to stress the need for other approaches in handling large quantities of scientific texts.

### *5.1 The index is pivotal*

In all literature on LB. the toughest nut to crack is the discrepancy between the unapproachability of the given index terms and the not entirely clear queries from the reader. The aim which IR research sets itself therefore is to merge the two different semantics. IR literature is described as an ongoing search for a renormalisation of the given index terms (mostly external identifiers) in useful formats for the searcher. In most operations it is understood that queries are

[3] As this work is oriented to the problems of large scale production and dissemination of scientific journal articles, the authors limit it to a first practical model which can be easily implemented in scientific publishing.

not necessarily answered by terms in the existing indexes. However, it is also taken as a starting point that proper manipulation of index terms will lead to the answer.
All IR research follows traditional indexing practice and starts with two implicit assumptions, namely: 1. the author represents the knowledge to be transferred correctly in the article (although sometimes professional indexers have to explicate it), and 2. in principle the reader knows what he or she is looking for (although sometimes professional librarians or information scientists have to rephrase the queries in the language of the database). Given these two implicit assumptions, the way is open for a formalisation of, firstly, the author's information into manageable quantities (see preceding section) and subsequently the information need of the reader into an equally manageable query.

In all IR models one tries hard to create formal mathematical methods to master the relation between the query and the existing index terms, whilst evaluation can only be made by people, who represent the informed user.

One has always to remember that in all models of IR, the correctness of the keyword index (or in so-called free text searching the author's own words) as the only available representation is taken for granted; the semantic components are considered central [23].

### *5.2 The limits off formal logic methods*

Belkin and Croft [24] in a review clearly indicate the weaknesses of existing (exact or partial match) retrieval methods. Appreciating the difficulties with manipulating simple index terms, in the fuzzy set approach one combines strict Boolean logic with partial relation of a document to a certain index term. 'In a fuzzy retrieval system an indexer instead of simply (sic!) assigning terms to a document, also indicates the strength with which he believes the term should be associated with the document' [25]. In probabilistic retrieval models, using statistical methods, probabilities are calculated that certain fixed index terms are relevant for the request.

Most models are purely mathematical methods based on information science theoretical approaches. The manipulation of lists of index terms becomes a mathematical skill, where, to a large extent, the practical use has faded away. It is telling that Boyce and Kraft [26] can conclude that 'we are unaware of a single theory in information science that suggests a testable phenomenon whose successful observation would add to its credence'. In an attempt to escape the self imposed harness of purely formal methods based on data retrieval and go 'beyond the mere keyword approach', van Rijsbergen [27, 28] even tries to formulate a non-classical logic for IR, as 'the main obstacles appeared to be an adequate computable model of meaning, and its use in information retrieval operations'. All this work is mounting evidence that IR is still in its infancy and badly needs new approaches.

### *5.3 'Need representation*'

It is striking to note that already in the classic introductions to IR research,

such as van Rijsbergen [11] or Salton and McGill [12], after elaborating all kinds of search techniques on online database systems, evaluation of the obtained results is by far the most difficult part.

To date no comprehensive models have been developed in which measurable quantities (like recall, precision, search time etc.) are treated satisfactorily. Grand old man Cleverdon [29] clearly represents librarians' frustrations as he correctly states: '.. . that the present-day services are, on the whole, inefficient and over-expensive products, packaged in the shiny wrapping of modem technology, and hostile to end-users...' He goes on to state that 'there has been the failure to realise and accept that retrieval of citations from a bibliographic database approximates to a random process'.

Recently Saracevic *et al.* [30-32] conducted a full scale study under as real-life a situation as possible to user contexts, of query formulation, search method and satisfaction with results. In this extremely elaborate study, not only has an attempt been made to analyse user requests, but also to distinguish between the different (cognitive) capacities of the searchers. The wealth of statistical results reported by the authors once again clearly proves the ambiguity of IR concepts. Many conclusions do not go beyond straightforward intuitive conclusions, if one already accepts that the actual database systems only direct the way in situations where the direction is already known in principle. Again, observations are confirmed that the degree of overlap in selection of search terms is small, as was the overlap in retrieved items. So, in other words, two people do not have the same thoughts from reading the same words.

This study splendidly underscores the thesis that IR, based on using keywords only, is unable to deal with the real problem of finding information one does not yet appreciate sufficiently to be able to phrase questions about in formalised terms.

The emerging option of having all articles in full text in the computer memory instead of bibliographic data and abstracts only will solve this problem only partially, as an external identifier might not direct to the most relevant part of the paper. Instead of taking the diffuse information need of the reader and the way he or she reads seriously, this is projected, or better, flattened onto already available knowledge, therewith constricting the whole operation to the transfer of known, codified information. Even if the reader with appropriate feedback systems is able to interact with the system by rephrasing, restricting, narrowing or broadening queries, the gist of the matter will never reach him from the retrieval-system.

## 6. A CALL FOR A RHETORIC SYNTAX

In the preceding pages a case has been built up against the idea that a semantic representation of knowledge in a scientific article can be constructed sufficiently correctly that an ignorant (but interested) reader will be able to use it for his or her knowledge gathering.

In all attempts to 'map' the contents of a paper in words, only the message

the author stresses is taken into account. The main objection against all this is that in all IR research it appears to be taken for granted that there must be some general and universal semantic mapping out there. As Warner [33] shows, hardly any interpenetration of IR and linguistic theory exists, underwriting Blair's [14] attempt to change the mood. Without challenging this idea further, the simple question arises as to why no more practical solutions and models are being investigated. After all, if a scientist is able to scan the written literature in print very rapidly, why then should it not be possible to use this ability as the starting point in information disclosure and retrieval?

Clearly the scientist is trained in scanning articles which have common characteristics in format. Interestingly enough, however, hardly anything has been written on this phenomenon except a first attempt by Bazerman [2]. Although the process of writing articles receives attention (see for example Latour [34], Latour and Woolgar [35]), hardly any work has been done in analysing the argumentational structure of the article as such. In a field such as law, it is accepted that the argumentational aspects are part and parcel of the discourse. In, for example, the works of Perelman and Olbrechts-Tyteca [6] and Perelman [36], extensive analyses are made on the aspects of persuasion and justification. For the natural sciences, and physics particularly, no serious studies of this kind exist. It looks as if the positivistic claim that science is proceeding slowly but certainly in the right direction is taken so seriously that even the structure of a scientific article resembles the formal logic of the mathematical tools used in the research. The standard paper is almost completely codified, and claims in its structure and language the same rigour of argumentation as does formal logic which is the backbone of the analyses of scientific results. Since a slow start with Gusfield [37] and his famous drinking driver in the eighties, a beginning has been made with the serious rhetorical analysis of scientific articles. Most of this work however is in the humanities [38] and only very little in science [39, chapters 6 and 9].

Following the SGML idea of having entity definitions without looking at the content (for example it is not important if the content of the entity 'surname' reads Frankenstein or B52) one can imagine that a structural representation, following the argumentation of the article, can be designed. On a meta-level a structural model, or to use a term of Woods [40], a 'structured taxonomy', of the lines of reasoning in scientific publications can be defined. Without underrating the importance of sociological and psychological research in establishing the actual style of writing, the first step into an argumentational analysis of scientific papers will be a phenomenological one. The fortunate situation is that the amount of material with which one could experiment is enormous and most of the material is not only in print, but also (although sometimes littered with all kinds of coding) available on magnetic tape from the publisher's typesetting facilities.

A quick first glance shows that the division of an article into sections and subheadings such as: Introduction, Experimental set-up, Design of the study, Theoretical excursion, Discussion, Conclusions, etc., claims more than it

suggests. Only if one assumes that the author is indeed as logical in his or her thinking as the structure of the article suggests, does a list of subheadings as suggested by Line [41] give some relief to the reader in his voyage of discovery. In almost every section assumptions are made, outside facts are mentioned or established points of view are adopted without discussion. In fact almost all sections contain elements of argumentation, some explicit, as they belong to the discussion of the subject and many implicit as they belong to the accepted body of knowledge within the research subject.

An interesting point in revealing the argumentational structure of a scientific discourse is that it ultimately might provide us with a different level of presentation of the work from that produced by a semantic network. The two representations are, to a large extent, complementary. One can imagine that a reader who wants to scan literature is only interested in retrieving those paragraphs or sentences of an article which deal with 'experimental conditions' or 'mathematical approximations' within a corpus of documents which is retrieved with a traditional keyword approach. In establishing a list of 'argumentational entities' such as 'references to own previous work' and 'references to other people's work', one can introduce this in the full text by coding in the same way as SGML.

Correlations between such entities can also be defined and, following the same philosophy as SGML, a presentation can be prepared using syntactic entities without going into depth of content or meaning analysis. The great advantage of an argumentational syntactic structure (or even template) is, that as no reference has to be made to content, the whole ambiguity of keyword assignment or changing semantics in the history of a subject is kept out. But not only are these types of ambiguities left aside. As argumentational entities are not necessarily composed of sequential (parts of) sentences or paragraphs, the original narrative style plays no role. In an experimental article for example, experimental constraints can be reported in the section dealing with the instrumentational setup as well as in the sections dealing with data collecting and data processing. In the argumentational entity 'experimental constraints' they are all put together. In the full (printed) version the article can be very concise or very chatty. As long as those elements which are of rhetorical value can be identified, the argumentation can be followed, and the assumptions, discussions, claims or results retrieved separately.

## 7. A TENTATIVE TEMPLATE FOR AN ARGUMENTATIONAL SYNTAX

In this section an attempt is made to provide a tentative template for an argumentational syntax in the case of an experimental physics article. To a certain extent, the listing of notions which structure the argumentation looks like a contents list. This is not so strange as, according to the many writers' aids, the conventional article claims that it takes the reader carefully by the hand and guides from simple premises logically to higher stages of knowledge and understanding.

The snag is however, and this is the crucial point, that the paragraphs of the

articles are small arguments in themselves, full of assumptions, new information, inferences etc. If a sentence reads 'following Einstein we can conclude that' it means that reference is made to other people's work, i.e. the work of Mr Einstein. Also, the conclusions of that work are adopted without discussion. The retrieval system will store it under 'references to other people's work'. The searchers/readers however can also only look at the conclusions section of the paper, and then decide that they want to know under what assumptions these concluding inferences are made. Therefore the retrieval system will be asked to show those sentences which belong to the notion 'theoretical assumptions', with the expected result that the above mentioned sample sentence will be retrieved. The sentence is linked to different notions.

As this simple example shows, no knowledge of the subject or of Mr Einstein and his theories is needed to declare the argumentational entities.

It is important in identifying the different argumentational notions (like the above mentioned 'theoretical assumptions') in texts that, in principle, there is no reason that their contents will not overlap. One and the same (part of a) sentence or paragraph can, for example, be listed as referring to 'previous own work' as well as to 'theoretical model'.

Another crucial feature is that very few rules between the argumentational notions are required. Contrary to formalised question-answering systems and also to some tendencies in argumentational theory to cast augmentations in formal structure, the goal here is only to represent the actual real life argumentational discourse. In that sense it is reasonable to speak of a phenomenological or pragmatic argumentational syntax. The few rules which, in a first approximation, can be designed are rules which connect mutually dependent notions. Therefore, the collection of (parts of) sentences throughout the paper which give other people's experimental results has a link with the full literature references to other experimental works. Another set of rules governs the relation of hierarchical notions, such as 'experimental constraints', which requires a notion 'experimental setup' (but not the other way around). Following the practice of the classic indexer one can think of defining relational operators of the kind *see* and *see also.*

In the table a tentative template is displayed which is in no way a claim for a full model, but serves only to help starting the discussion on a possible argumentational syntax. In the table various notions which might serve as entities for an argumentational analysis of scientific articles are given. A single number indicates that this notion can be considered as central, while a multiple number means that the broader notion has been defined. In a way they serve as a *see* operator in the sense that it is compelling or in other words; there exists a hierarchical link. After every definition of the first four categories also some references to other numbers of the list are given.

These can be seen as a kind of *see also* operators in the sense that it is suggested that one also look under the entry mentioned (except to categories 10 and 11 which are obviously connected to every other entry), in other words, non-hierarchical links. As this table is only given to explicate the basic idea, no more than illustrative value can be attached to it, while completeness is far from being claimed.

1. Definition of the research subject in broad terms
   1.1 Redefinition of the problem in the actual research context (2.1/2.2/ *5.* 1/ 5.2)
2. Experimental setup (1.1)
   2.1 Experimental constraints
   2.2 Experimental assumptions *(5.2)*
   2.3 Experimental ambiguities (4)
   2.4 Relation of experimental setup with other experiments
3. Data collection
   3.1 Data handling methods (5.2)
   3.2 Data handling criteria (2.1)
   3.3 Error analysis (2.3)
4. Presentation of raw experimental data (2.1/2.2)
   4.1 Presentation of smoothed experimental data (2.2/3.3/5)
   4.2 Pointers to pictorial or tabular presentation
   4.3 Comparison of own data with other results (2.4)
5. Theoretical model
   5.1 Theoretical constraints
   5.2 Theoretical assumptions
   5.3 Theoretical ambiguities
   5.4 Relation of theoretical elaboration with other works
6. Theoretical/mathematical elaboration
7. Presentation of theoretical results/predictions
   7.1 Comparison with other theoretical results
   7.2 Pointers to pictorial or tabular presentation
8. Comparison of experimental results with own theoretical results
   8.1 Comparison of experimental results with other theoretical results
   8.2 Pointers to pictorial or tabular presentation
9. Conclusions
   9.1 Experimental conclusions
   9.2 Theoretical conclusions
10. Reference to own previous published work
   10.1 Reference to own work in progress
11. Reference to other people's published work
   11.1 Reference to other people's work in progress

## 8. A POSSIBLE TECHNICAL IMPLEMENTATION

It is always easier to plead for a new approach than to pursue it practically. In this article a call for a new type of grammar in representing scientific articles is made; so that *noblesse oblige* in this paragraph therefore a practical solution for an implementation is proposed.

For quite some years in the computer industry, experiments have already been under way with so-called hypertext. Hypertext comprises all attempts to link the contents of fields in a database with well defined areas (windows) on a computer screen.

An excellent overview of the essence, advantages and disadvantages of hypertext is given by Conklin [42]. For an enthusiastic review of possibilities see Davenport and Cronin [43] and references therein. As Conklin formulates it, hypertext is the ability to perform high speed branching transactions on textual chunks. Thus, different parts of a text can be linked in all possible ways which the system designer or user wants. Hypertext is essentially a database method of organising information, as well as being a representation scheme. Within hypertext, links between parts of texts are defined. So, for example, within a running text the option of making a jump to another text, figure or graph can be made. The definition of 'links' between different text (or figure) chunks is the core of a hypertext approach. The easiest way to imagine hypertext is to think of an encyclopedia.

Suppose the reader looks up the term 'grammar'. The computer shows the text of the entry on the screen. After some lines the name 'De Saussure' appears. Within hypertext this name is then linked to the biography of the same person. When one arrives at this point., one can be linked further to the map of his town of birth etc. Within hypertext, one can distinguish non-hierarchical referential links, which connect points or regions in the text, as well as hierarchical links.

Hypertext can be excellently used in our context. Comparing the similarity to semantic networks, Conklin points out: 'Hypertext nodes can be thought of as representing single concepts or ideas, internode links as representing the semantic interdependencies among these ideas, and the process of building a hypertext network as a kind of informal knowledge engineering. The difference (with Artificial Intelligence) is that AI knowledge engineers are usually striving to build representations which can be mechanically interpreted, whereas the goal of the hypertext writer is often to capture an interwoven collection of ideas without regard to their machine interpretability'. From this quotation it is clear that the developers see hypertext as a relational database where the entities can be freely defined and linked. In this sense hypertext is the most open semantic network. Limiting hypertext to text objects seen only as semantic entities (with hierarchical and non-hierarchical links) would reduce the method to a sophisticated thesaurus structure, however not with nouns or verbs, but with whole text chunks as building blocks.

A hypertext environment tailored to our problem can be more than a very open and flexible non-linear representation of text. Such an approach will be characterised by the following points:

1. We do not use text parts themselves as elements in the network, but the argumentational notions which contain parts of the text. In a sense we propose to use the argumentational meta-level as input for a hypertext structure.

2. An SGML-like grammar has to be developed which links the entities, next to a thesaurus type of representation. The advantage is that the user who is not a specialist can then easily go back and forth between conceptual entities (within

the structure of the argumentation) and semantic (meaning) entities which link words to other words or external identifiers.

3. The chunks of text are not fully independent in the sense that they represent nodes in a network. Some parts of sentences or paragraphs can show up in different nodes as they contribute to a differentiation of different notions. On the other hand a full sequential representation is always available in order to read the document as it is published in print.

4. As many articles will be loaded in such a system, text 'chunks' of very many different articles, with the same 'argumentational tag' (for example experimental constraints) can be retrieved together. Subsequently this set of texts (for example all relevant sentences, grouped per article, in relation to experimental constraints) can be read and judged in order to decide how to go on.

In conclusion we can say that hypertext allows us to develop a multi-layered representation of a scientific article. One layer represents the article in the full form as it is published, another layer represents the work in terms of internal, external, and transmittal identifiers, while again another layer represents the document in its argumentational structure. It will be clear that no hierarchical order in these various layers can be established. Although the original article as provided by an author is the source of everything, its use by a reader is no longer confined to this type of representation of scientific results.

Further investigations along the lines described are badly needed in our scientific world where information growth has not been matched by correspondingly improved capacity to penetrate the stored knowledge.

It is to be hoped that this article contributes to the implementation of a research programme addressing these issues.

## ACKNOWLEDGEMENTS

The critical and very valuable comments of Antje Melissen, Arie Manten, Arie de Ruiter, Barrire Stern, Jan Bleeker, Tor Henriksen and Karen Hunter are gratefully acknowledged. Without the invaluable technical help of Bella Goossens and Janette Young this article would never have been finished.

## REFERENCES

1. CLYNE, M. Cultural differences in the organization of academic texts. *Journal of Pragmatics, 11,* 1987, 21 1—247.
2. BAZERMAN. C. Physicists reading physics, schema-laden purposes and purpose-laden schema. *Written Communication,* 2(l), 1985, 3—23.
3. ELLIS. D. A behavioural approach to information retrieval system design. *Journal of Documentation. 45(3),* 1989, 17 1—212.
4. ELLIS, D. A behavioural model for information retrieval system design. *Journal of Information Science, 15(4),* 1989, 237—247.

5. MULUNS, N., SNIZEK, W. *and* OEHLER., K. The structural analysis of a scientific paper. *In:* VAN RAAN, A. F. J.*, ed. Handbook of quantitative studies of science and technology.* Amsterdam: North-Holland, 1988, 81—106.
6. PERELMAN, C. *and* OLBRECHTS-TYTECA, L *The new rhetoric, a treatise on argumentation.* Notre Dame, IN: University of Notre Dame Press, 1969.
7. DAY, R. *A. How to write and publish a scientific article.* Philadelphia, PA: ISI Press, 1979. .. -
8. GATES, V. Stuperspace. *Physica, 15D(l &* 2), 1985, 289—293.
9. PINCH, T. Towards an analysis of scientific observation: the externality and evidential significance of observational reports in physics. *Social Studies of Science, 15.* 1985, 3—36.
10. BAZERMAN, C. Modem evolution of the experimental report in physics; spectroscopic articles in *Physical Review,* 1893—1980. Social Studies of Science, 14. 1984, 163—196.
11.VAN RIJSBERGEN, C.J.. *Information retrieval.* 2nd ed. London: Butterworth, 1979.
12. SALTON, G. *and* MCGILL M. J. *Introduction to modern information retrieval,* London: McGraw-Hill, 1983.
13. B1ESRICHER., P., FUHR, N., LUSTIG, G., SCHTWANTNER, M. *and* KNORZ, G. The automatic indexing system AIRPHYS: from research to application. *In:* CHIARAMELLA, Y., *ed. Proceedings of the 1988 ACM Conference on Research and Development in Information Retrieval, Grenoble, 1988.* New York: ACM, 1988, 333—342.
14. BLAIR, D.C. *Language and representation* in *information retrieval.* Amsterdam: Elsevier, 1990.
15. LIUDDY, E., BONZI, S., KATZER, J. *and* ODDY, E. A study of discourse anaphora in scientific abstracts. *Journal of the American Society for Information Science. 38(4),* 1987, 255—261.
16. LIDDY, E. D. Anaphora in natural language processing and information retrieval. *Information Processing and Management, 26(l),* 1990, 39—52.
17. MACROBERTS, M. H. *and* MACROBERTS, B. R. Problems of citation analysis: a critical review. *Journal of the American Society for Information Science. 40(5),* 1989, 342—349.
18. BARRON, D. Why use SGML. *Electronic Publishing, 2(1),* 1989, 3—24.
19. INTERNATIONAL STANDARDS ORGANIZATION. *1S0 8879. Information processing — text and office systems — standard generalized markup language (SGML).* Geneva: ISO, 1986.
20. INTERNATIONAL STANDARDS ORGANIZATION. ISO/IEC TR *9573. Information processing. SGML support facilities techniques for using SGML.* Geneva: ISO/IEC, 1988.
21. BRYAN, M. *SGML: an author's guide to the standard generalized markup language.* Wokingham, UK: Addison-Wesley, 1988.
22. KIRCZ, .J.G.. *and* BLEEKER, J. The use of relational databases for electronic and conventional scientific publishing. *Journal of Information Science, 13(2),* 1987, 75—89.
23. SALTON, G. Another look at automatic text-retrieval systems. *Communications of the ACM, 29(7),* 1986, 648—656.
24. BELKIN, N.J. *and* CROFT, W. B. Retrieval techniques. *In:* WILLIAMS, M. E., *ed. Annual review of information science and technology.* Vol. 22. Amsterdam: Elsevier Science Publishers, 1987, 109-145.
25. BOOKSTEIN, A. Probability and fuzzy-set applications to information retrieval. *In:* WILLIAMS, M. E., *ed. Annual review of information science and technology.* Vol. 20. Knowledge Industry Publ.. Inc., 1985, 117-151.

26. BOYCE. B. R. *and* KRAFT, D. H. Principles and theories in information science. *In:* WILLIAMS, M. E., *ed. Annual review of information science and technology.* Vol. 20. Knowledge Industry Publ. Inc., *1985, 153-178.*
27. VAN RIJSBERGEN, C.J. A new theoretical framework for information retrieval. *SIGIR Forum, 21(1—2),* Fall-Winter 1986-1987, 23-29.
28. VAN RIJSBERGEN C.J. A non-classical logic for information retrieval. *The Computer Journal. 29(6),* 1986, 481-485.
29. CLEVERDON, C. Optimizing convenient online access to bibliographic databases. *Information Services & Use. 4,* 1984, 37-47.
30. SARACEVIC, T. KANTOR, P., CHAM1S. A. Y. *and* TRIVISON. D. A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science, 39(3),* 1988, 161-176.
31. SARACEVIC, T. and KANTOR. P. A study of information seeking and retrieving. II. Users, questions, and effectiveness. *Journal of the American Society for Information Science. 39(3),* 1988, 177-196.
32. SARACEVIC. T. *and* KANTOR, P. A study of information seeking and retrieving. III. Searchers, searches, and overlap. *Journal of the American Society for Information Science, 39(3),* 1988, 197-216.
33. WARNER, A. J. Quantitative and qualitative assessments of the impact of linguistic theory on information science. *Journal of the American Society for Information Science, 42(l),* 1991, 64-71.
34. LATOUR, B. *Science in action.* Milton Keynes: Open University Press, 1987.
*35.* LATOUR, B. *and* WOOLGAR, S. *Laboratory Life.* 2nd ed. Princeton, NJ: Princeton University Press, 1986.
36. PERELMAN, C. *Retorica en argumentatie.* Baarn: Basisboeken Ambo, 1979. (Dutch translation of *L 'Empire rhétorique et argumentation,* 1977).
37. GUSFIELD, J. The literary rhetoric of science: comedy and pathos in drinking driver research. *American Sociological Review, 41* (February), 1976, 16-34.
38. SIMONS, H. W. *The rhetorical turn.* Chicago: University of Chicago Press, 1990.
39. GROSS. A. G. *The rhetoric of science.* Cambridge, Mass.: Harvard University Press, 1990.
40. WOODS, W. A. Important issues in knowledge representation. *Proceedings of the IEEE.* 74(10), 1986, 1322-1334.
41. LINE, M. B. Redesigning journal articles for on-line viewing. *In:* LINE, M. B. *Lines of thought: selected papers of Maurice B. Line.* London: Clive Bingley, 1988.
42. CONKLIN, J. Hypertext: an introduction and survey. *Computer, 20* (September) 1987, 17-41.
43. DAVENPORT, E. *and* CRONIN, B. Hypertext and the conduct of science. *Journal of Documentation,46(3),* 1990, 175-192.

*(Revised version received 15 April 1991)*