TOWARD A CATALOGUE OF CITATION-RELATED RHETORICAL CUES IN SCIENTIFIC TEXTS

CHRYSANNE DI MARCO AND ROBERT E. MERCER
University of Waterloo, Waterloo, Ontario, Canada
and
The University of Western Ontario, London, Ontario, Canada

Scientific citations establish an explicit network of relationships among mutually relevant articles within a research field. By convention, authors include citations in their papers to indicate works that are foundational in their field, background for their own work, or representative of complementary or contradictory research. But, determining *a posteriori* the nature of the exact relationship that an author intended between a citing and cited paper is often difficult to ascertain. To address this problem, the aim of formal citation analysis has been to categorize and, ultimately, automatically classify scientific citations. In previous work, Garzone and Mercer (2000) presented a system for citation classification that relied on characteristic syntactic structure to determine citation category. In this present work, we extend this idea to propose a more general catalogue of stylistic and rhetorical techniques that may provide just such an appropriate basis for categorization.

Key words: discourse analysis, text understanding, rhetoric of science

1. THE CITATION PROBLEM: AUTOMATING CLASSIFICATION

1.1. Motivation for the research

A *citation* may be formally defined as a portion of a sentence in a citing document which references another document or a set of other documents collectively. For example, in sentence 1 below, there are two citations: the first citation is *Although the 3-D structure... progress*, with the set of references (Eger et al., 1994; Kelly, 1994); the second citation is *it was shown... submasses* with the single reference (Coughlan et al., 1986).

(1) Although the 3-D structure analysis by x-ray crystallography is still in progress (Eger et al., 1994; Kelly, 1994), it was shown by electron microscopy that XO consists of three submasses (Coughlan et al., 1986).

A *citation index* is used to enable efficient retrieval of documents from a large collection—a citation index consists of source items and their corresponding lists of bibliographic descriptions of citing works. The primary purpose of scientific citation indexing is to provide researchers with a means of tracing the historical evolution of their field and staying current with on-going results. However, with the huge amount of scientific literature available, and the growing number of digital libraries, standard citation indexes are no longer adequate for providing precise and accurate information. Too many documents may be retrieved in a citation search to be of any practical use. And, filtering the documents retrieved may require great effort and reliance on subjective judgement for the average researcher. What is needed is a means of better judging the relevancy of related papers to a researcher's specific needs so that only those articles most related to the task at hand will be retrieved. For this reason, the goal of *classifying* citations evolved out of citation analysis studies. If, for example, a researcher is new to a field, then he may need only the foundational work in the area. Or, if someone is developing a new scientific procedure, he will wish to find prior research dealing with similar types of procedures.

A key factor in enhancing the quality of a search through related documents will be the ability to indicate the nature of the citation relationships that are of interest, which, in turn, is directly related to the comprehensiveness (coverage and granularity) of the citation classification scheme. A trade-off exists, therefore, between accuracy and usefulness of results and the amount of effort required to obtain this degree of precision—the larger the number of categories and the finer-grained the classification scheme, the more difficult it will

be to pin down the exact linguistic cues in the citing article that indicate why those categories are being used.

In earlier work, Garzone and Mercer[1] ([Garzone1996], [Garzone and Mercer2000]) proposed a citation classification scheme that, with 35 categories, was both more comprehensive than the union of all of the previous schemes and also amenable to implementation in an automated citation classifier. The approach taken was to search for structural cues in citing sentences that could be matched against a *pragmatic grammar* consisting of 195 lexical matching rules and 14 parsing rules to classify citations according to a citation's cue words and location in the article. The automated citation classifier was evaluated on a set of biochemistry and physics articles, with resulting fair performance on previously unseen and good performance on previously seen articles. We now propose to extend this idea by using a variety of rhetorical cues within citation sentences and the surrounding text as a stylistic basis for categorization.

1.2. Background to the research

As Garzone and Mercer ([Garzone1996], [Garzone and Mercer2000]) demonstrated, the problem of classifying citation contexts can be based on the recognition of certain *cue words* or specific word usages in citing sentences. For example, in sentence 1, the phrase *still in progress* may be taken to indicate that the citation is referring to work of a concurrent nature. As well, the use of the past tense of the verb in the phrase *was shown* indicates that a key result is discussed in this previous work.

In order to recognize these kinds of cue-word structures, Garzone and Mercer based their classifier system on what they called the *pragmatic parser*. The knowledge used by the parser to determine whether a certain pattern of cue words has been found was represented in a *pragmatic grammar*. As Garzone and Mercer explain: "Our choice of the term 'pragmatic grammar' (and hence 'pragmatic parser') has been motivated by the existence of semantic grammars where specialized lexical categories are based on their semantic properties. Some constituent categories have been motivated by the *function* of the constituent in this particular domain of citation classification in scientific journals. The purpose of the pragmatic grammar is to suggest the function of a citation."

The purpose of the grammar was to represent the characteristic structural patterns that corresponded to the various citation functions (i.e., categories) in their classification scheme. The grammar was developed by manually extracting and studying citations from a set of journal articles (8 physics and 6 biochemistry). The rules in the grammar were of two types: lexical rules based on cue words which were associated with functional properties and grammar-like rules which allowed more sophisticated patterns to be associated with functional properties.

For our present purposes, the nature of the cue-word rules is most relevant. As an example, the grammar contained a rule specifying that if any of the cue words *postulated*, *reads*, or *reported* were found in the Results section of the journal article, the word's presence would indicate that the citation should be classified under the category *used for developing new hypothesis or model*. As we noted earlier, 195 such lexical matching rules were constructed. The success obtained by Garzone and Mercer from using this cue-word–based approach for their classifier suggested that there may be value in looking for a more systematic and general definition of cues based on a document's rhetorical structure. An additional outcome of Garzone's experiment that seems noteworthy to pursue was the recognition of the important role that the preceding and following sentences could play in determining the category of a

[1] We use some definitional material from Garzone and Mercer (2000) with permission.

citation. Clearly, it seems useful to investigate whether incorporating some form of discourse analysis may enhance the current state of automated citation classifiers.

2. THE ROLE OF DISCOURSE STRUCTURE IN CITATION ANALYSIS

2.1. Our approach: Using fine-grained rhetorical information in citation analysis

We take as our starting-point the premise that knowing the fine-grained rhetorical structure of a scientific article can help tremendously in citation classification. We base this premise on two arguments: the well-established body of work in rhetorical theory may be used in analyzing the global structure of scientific discourse (e.g., [Fahnestock1999], [Gross1996], [Myers1991]), and more-recent studies have demonstrated the role of fine-grained discourse cues in the rhetorical analysis of general text. We intend to show that this latter work, as exemplified by Knott [Knott1996] and Marcu [Marcu1997], may, together with models of scientific argumentation, provide a means of constructing a systematic analysis of the role citations play in maintaining a network of rhetorical relationships among scientific documents.

In the long-term, our intention is to show that there is a direct mapping from the fine-grained argumentation structure of scientific discourse to formal rhetorical relations that express the communicative purpose of the context within which they are used. It is our contention that citations are a key part of the fine-grained rhetorical structure of a scientific argument, acting as contextually motivated items to help construct the very nature of the argument. As such, it should be possible to show that citations can be mapped to the local rhetorical relations that underlie the scientific discourse structure. These rhetorical relations in turn can assist in classifying a citation by providing an explanation of the author's purpose in using the citation to link to a certain article. As a first step then, we need to show that there are indeed overt structural cues in scientific discourse that can be detected by automated means, that these are types of cues that may be associated with rhetorical relations, and that such cues play a significant role in citation contexts.

2.2. Background: Cue phrases in discourse analysis

Knott: Defining a 'Cue Phrase'. In the most basic sense, a *cue phrase* can be thought of as a linguistic conjunction or connective that assists in building the coherence and cohesion of a text. For example, in passage 2, the use of *However* may be taken as an indication that there is some kind of semantic relationship between the two sentences—in this case, the second sentence provides a contrast to the first.

(2) I wanted to go outside today. However, it was so cold that I decided to stay home and read instead.

Various more-formal definitions of a cue phrase exist, and Knott [Knott1996] lists several of these: "For instance, Cohen (1984) defines 'clue words' as 'special words or phrases directly indicating the structure of the argument to the hearer'; Hirschberg and Litman (1993) define cue phrases as 'words and phrases that directly signal the structure of a discourse'." But, as Knott adds, such definitions already require that one knows the structure of the discourse so that the definition is circular. As an alternative and more-formal definition, Knott proposed a precise test for cue phrases that he then used in analyzing academic texts to construct a corpus of cue phrases. This corpus was later enlarged by Marcu [Marcu1997], and is the one that we use in our own studies.

In developing his corpus of cue phrases, Knott used the following classification of cue phrases into five syntactic groups (pp. 66–67), a classification we will extend below:

Coordinators: These cue phrases always appear in-between the clauses they link; the clauses can be in separate sentences or in the same sentence. For example:

(3) An object may move but it remains the same object.

Subordinators: These introduce subordinate clauses in complex sentences. For example:

(4) Although it is common sense that labels are related, this is a difficult idea to explicate.

Conjunct adverbs: These modify whole clauses, and can appear at different points within them, although there is often a default position for particular phrases. For example:

(5) We will select only those hypotheses we deem relevant. As a consequence, our discussion differs from the usual views.

Prepositional phrases: These often contain propositional anaphora referring back to the previous clause. For example:

(6) It has a high degree of opacity. In that respect it resembles glass.

Phrases which take sentential complements: These often introduce a particular intentional stance with respect to the content of the clause they introduce. For example:

(7) It may seem that we are making too much of orientation; but characteristic orientation is not an idiosyncrasy.

In addition to providing a formal means of defining cue phrases and compiling a large catalogue of phrases (over 350), Knott's other main result is of particular significance to us: he combines the two methods hitherto used in associating cue phrases with rhetorical relations to argue that "cue phrases can be taken as evidence for relations precisely if they are thought of as modelling psychological constructs" (p. 22). For our purposes then, Knott's supporting demonstration for this argument allows us to rely on his result that there is indeed a sound foundation for linking cue phrases with rhetorical relations.

Marcu: Formalizing Rhetorical Relations. A necessary requirement for our hypothesis that citation classification can be based on the analysis of detailed rhetorical structure is that such rhetorical information may be obtained through automated means. Many types of rhetorical relations have been proposed, from a minimal set of purely coherence relations to extensive lists of more pragmatics-based relations involving the communicative purpose of a text. For our intended citation analyses, the pragmatic type of rhetorical relation is most applicable, and, of these, Rhetorical Structure Theory (RST) [Mann and Thompson1988] provides the current most popular set of rhetorical relations for use in Computational Linguistics. Marcu [Marcu1997] extended the work on RST in several ways that are key to our purposes: he gave a formalization of RST; a *rhetorical parsing algorithm* for deriving the valid discourse structure of unrestricted texts; and, most importantly, an implementation of this algorithm in the form of a *rhetorical parser*. Furthermore, the rhetorical parser uses cue phrases in order to "hypothesize rhetorical relations between clause-like units, sentences, and paragraphs…" (p. 142). The existence of such a rhetorical parser fulfils our requirement that the analysis of rhetorical relations may be automated, and we plan to investigate the use of Marcu's parser in our later work.

3. A CATALOGUE OF STYLISTIC AND RHETORICAL METHODS TO INDICATE CITATION COHERENCE

The underlying premise of studies on the role of cue phrases in discourse structure (e.g., [Halliday and Hasan1976], [Knott1996], [Marcu1997]) is that cue phrases are purposely used by the writer to make text coherent and cohesive. With this in mind, we are analyzing a dataset of scholarly science[2] articles. Our current task is to begin to catalogue the fine-grained discourse cues that exist in citation contexts. The first stages of this catalogue are presented in the next sections. Our initial analysis confirms that authors have a rich set of linguistic and non-linguistic methods to establish discourse cues in citation contexts.

Description of the Analysis. We are using a dataset of 51 scholarly science articles. Most of these articles are written in the IMRaD[3] style or a minor variant of that style. Previously, we analyzed the frequency of the cue phrases from [Marcu1997] in these articles. We have reported strong evidence that these cue phrases are used in the citation sentences and the surrounding text with the same frequency as in the article as a whole ([Mercer and DiMarco2003]).

After performing this initial frequency analysis, we have begun to catalogue other discourse and stylistic aspects of these articles that strongly indicate rhetorical relations. We are currently interested in those relations that provide evidence of the citation category. This cataloguing is work in progress. We are constantly fine-tuning our procedures since we are always encountering nuances in the way the data expresses the rhetorical structure. We report below on the current state of this catalogue.

Our current catalogue. In addition to the cue phrases suggested by Knott ([Knott1996]), we consider the following items which result from our focus on scientific articles and our concern with the inter-article connections provided by citations.

Several types of syntactic stylistic usage provide rhetorical contexts that may serve to indicate the nature of the citation. For example, the use of syntactic symmetry or parallelism can act as a cue for one or more citations.

1. Symmetry.

   (8) The values. . . are in good agreement with. . . as well as. . . (ref).

   The comparison set up by *as well as* indicates the symmetric structure of the citation sentence in which the positive (or negative) polarity of the first half of the comparison context is maintained in the second half. In this case it is positive, so the (ref) in the second half will refer to a positive situation as well.

   (9) . . . it is still an open question whether. . . or. . . (ref, ref).

   In this citation sentence, the use of *whether. . . or* creates a specific kind of symmetry, parallelism in the form of matching rhetorical contexts, which present us with two plausible alternatives. The sentence ends appropriately with two references. We know also that for these alternatives to be considered seriously, there must be evidence in the scientific literature to support each of them. Although it may be the case that each reference discusses both alternatives, the more likely scenario, given the scope of scientific articles, is that the first reference has evidence for the first alternative and the second has evidence for the second alternative.

[2] We are currently working with two scientific genres, biochemistry and physics.
[3] Introduction, Method, Results, and Discussion.

2. Repetition (words/phrases).

   (10) We found that the temperature shift of the TA branch is due to an exchange of eigenvectors.... [This is discussing the work reported here on $KNBO_3$.] We note that an analogous exchange of eigenvectors may cause the anomalies reported in the TA branch of $SrTiO_3$ (ref).

   Repetition of words and phrases may be considered a form of lexical 'parallelism'. In this example, the citation refers to a paper that discussed (*reported*) similar work on a different compound ($SrTiO_3$ versus $KNBO_3$) that has analogous results which could account for the repeated phrase.

Other forms of rhetorical cueing rely on more-general aspects of a citation sentence's structure or a specific citation placement to indicate the nature of the citation.

3. Use of lists (temporal, examples of, etc.).

   (11) [The preceding sentence contains a temporal list indicating three phase transitions that depend on decreasing temperature.]

   All these transitions are connected to a continuous phonon softening with decreasing temperature (ref).

   The list structure used in this citation context is significant because it sets up the topic which is then referred to in the following sentence that contains the citation. The deictic *these* in the citation sentence refers to the common theme (*phase transitions*) given by the list. [Also note that there is a repetition of *decreasing temperature*: see above for discussion of lexical repetition.] Knowing that *phonon* is in the article's title, it is reasonable to conclude that (ref) discusses the connection among phase transitions, phonon softening, and decreasing temperature.

   (12) Identification of such catalytic residues has been attempted in a number of enzymes using chemical modification (ref), x-ray crystallography (ref), or homology searches coupled with site-directed mutagenesis of residues thought to be involved (ref).

   Very frequently, list structures in the form of enumerations are accompanied by citations associated with each element of the list. These types of citation sentences may also be marked explicitly by use of such terms as *a number of* (as above), *respectively*, or *as follows*.

4. Citation placement.

   (13) Cycles of energy minimization and refinement of temperature factors (ref) were indispersed with inspection of difference density and [technical jargon] maps.

   Some forms of stylistic usage specific to citation sentences ('citationese') appear to be associated with the specific placement of citations within the sentence. In this sentence, the citation is located immediately following the subject noun phrase (NP), as opposed to being at the end of sentence, say. Since (ref) is attached to the subject NP, the paper it cites obviously discusses that topic and not the topic of the main verb or its complement or any connections between these syntactic elements.

A number of citation cues are lexically related, whether by specific morphological usage (both general and domain-specific) or through exact lexical choice.

5. Lexical morphology.

   (14) Unlike the receptors. . . in which. . . is intracellular, in neu it is extracellular. . . (ref).

   Knowing that the use of the morphemes *intra* and *extra* sets up a binary contrast, we can further hypothesize that *unlike* is providing a binary contrastive context. Furthermore, the domain of contrast is now apparent: the cell interior/exterior.

6. Domain morphology.

   (15) Pure $KTaO_3$ is. . . . [The sentence following contains the phrases *stabilize the soft phonon* and *suppress the. . . phase transition* as well as a citation reference (ref).]

   Pure $KNbO_3$, on the other hand,. . . . [The sentence following contains the phrase *continuous phonon softening* and a citation reference.]

   In the previous discussion we have the introduction of the symbol *KTN* which makes a connection between these two compounds, and, importantly, *KTN* is used in the article's title. These two contexts show a contrasting result (*on the other hand*) for these two compounds which both belong in a domain of interest to the article. This will be important background information for the article.

7. Lexical choice.

   (16) The relative impact on activity of mutations in the catalytic acid and the catalytic nucleophile was unexpected and not in line with commonly observed effects (ref), (ref).

   Many occurrences of citations appear to be used when a negative result (*not in line*, as above) or unusually low level of a substance, activity, etc., is noted (e.g., *no glucanase activity*, *activity is. . . very low*,

   (17) . . . this may be why the three least probable side-by-side residue pairs involving amides in proteins are [list of pairs] (ref).

   (18) . . . a PDGF receptor mutant with a truncation of 141 amino acids from the C terminus displays almost no kinase activity, whereas a mutant with a deletion of 98 amino acids retains its kinase activity (ref).

   It is significant, we feel, that many 'negative' situations (and associated citations) in scientific writing appear to be marked by lexical means, both overt and more subtle. In the two sample citation sentences above, the words *least probable*, *truncation*, and *deletion* are 'hidden' indicators of negative results of some kind, along with the explicit cue, *almost no kinase activity*.

   (19) The enzymic hydrolysis of polysaccharides and glycosides is a critically important process in the metabolism of plants, animals, and microorganisms (ref).

   We have also observed that often many 'extreme' situations, both positive and negative, cue for citations. The example above is marked by the occurrence of the signalling phrase, *critically important*; other examples we have found include *almost totally (buried)*, *near complete loss*, *high levels*, and *major role*.

A final group of citation cues characterize the kind of rhetorical mannerisms used in scientific writing: procedural terms, scales, and 'reporting' style, for example.

8. Procedures.

   (20) Total RNA was prepared by...extraction (ref), separated on...using standard procedures (ref), and transferred to....

   Scientific procedures are sequences of steps and often described in list format in scientific papers. The verb used in each of the list items thus tends to have a strong procedural sense. In this example, the word *procedures* is even used in one of the list items. Citations are characteristically associated with the names of procedures to indicate the sources in which the procedure is described in more detail.

9. Scales.

   (21) [In a sequence of sentences, there is a discussion of phase transitions for different ranges of concentration of niobium.]

   Concentration is on the scale 0%–100%. The discussion referred to in the example above divides the scale into ranges. Each range has a different phase transition associated with it. One of the sentences uses the term *dilute*, which is also a meaningful term for the *concentration* scale. The concentrations of 0.8% and 0.9% are discussed at some length as they have importance for the remainder of the scientific article. These numbers are used throughout the article. One of the samples studied by the authors has a 0.8% concentration of niobium. The interpretation of various contexts in the article depend on knowing that 0.8% and 0.9% have this importance.

10. Reporting.
    Many examples of 'reporting' style appear to be used to signal a citation, in effect, making reference to the historical record of the author's own or other works. These cues often take forms similar to *It should be mentioned that...*, *It has been previously suggested that...*, and *As expected from previous studies...*. In the first two instances, the cues seem to belong to Knott's category of cue phrases which take sentential complements, i.e., to introduce a particular stance (here a reporting stance) with respect to the content of the clause they set up for.

4. CONCLUSIONS AND FUTURE WORK

We have set out to catalogue the rich variety of stylistic and rhetorical cues that authors use to create intra-textual and inter-textual coherence in scholarly scientific writing. Our analysis of 51 scholarly science articles indicates that there is an extremely rich set of discourse cues in scientific writing and citation passages. Our initial foray into the use of discourse cues to signal coherence with cited material has suggested a number of exciting possibilities. Knott ([Knott1996]) has suggested two categories—propositional anaphora and sentential complements that introduce an intentional stance—that appear to be used quite frequently in citation style. We have begun to catalogue other types of 'citationese' cues specific to the genre of scientific writing, cues specific to the domain of the article, and cues correlated with stylistic structure (e.g., lists, type of sentence opening).

In our cataloguing of such cues, we are continuing to develop two themes in the research. The first is the use of the rhetoric of science, specifically, models of scientific argumentation, to provide a basis for the detection of cues in citation contexts. The second theme is the development of several test corpora with pre-classified citations that may then be used in

evaluating our automated system. These corpora are being constructed in various ways, for example, random selection from a single genre, papers from different genres, papers with co-citations, and papers by the same author.

Our ultimate purpose in developing the catalogue is to identify linguistic cues that may be used as a means of determining the function of citations. Based on Knott, Marcu, and others, we can expect to be able to associate cue phrases with rhetorical relations as determiners of citation function. The interesting question then becomes: can we extend textual coherence/rhetorical relations signalled by cue phrases to extra-textual coherence relations linking citing and cited papers?

REFERENCES

[Fahnestock1999] J. Fahnestock. 1999. *Rhetorical Figures in Science*. Oxford University Press.

[Garzone and Mercer2000] M. Garzone and R.E. Mercer. 2000. Towards an automated citation classifier. In *Proceedings of the 13th Biennial Conference of the CSCSI/SCEIO (AI'2000)*, pages 337–346. Lecture Notes in Artificial Intelligence, volume 1822, H.J. Hamilton (ed.), Springer-Verlag.

[Garzone1996] M. Garzone. 1996. Automated classification of citations using linguistic semantic grammars. M.Sc. Thesis, The University of Western Ontario.

[Gross1996] A.G. Gross. 1996. *The Rhetoric of Science*. Harvard University Press.

[Halliday and Hasan1976] M.A.K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman Group Limited.

[Knott1996] A. Knott. 1996. A data-driven methodology for motivating a set of coherence relations. Ph.D. thesis, University of Edinburgh.

[Mann and Thompson1988] W.C. Mann and S.A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3).

[Marcu1997] D. Marcu. 1997. The rhetorical parsing, summarization, and generation of natural language texts. Ph.D. thesis, University of Toronto.

[Mercer and DiMarco2003] R.E. Mercer and C. DiMarco. 2003. The importance of fine-grained cue phrases in scientific citations. In *Proceedings of the 16th Conference of the CSCSI/SCEIO (AI'2003)*, page (to appear), Halifax, Canada.

[Myers1991] G. Myers. 1991. *Writing Biology*. University of Wisconsin Press.