The importance of fine-grained cue phrases in scientific citations

Robert E. Mercer[1] and Chrysanne Di Marco[2]

[1] University of Western Ontario, London, Ontario, N6A 5B7,
mercer@csd.uwo.ca
[2] University of Waterloo, Waterloo, Ontario, N2L 3G1,
cdimarco@uwaterloo.ca

Abstract. Scientific citations play a crucial role in maintaining the network of relationships among mutually relevant articles within a research field. Customarily, 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 the nature of the exact relationship 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 that fine-grained cue phrases within citation sentences may provide just such a stylistic basis for categorization.

1 The Citation Problem: Automating Classification

1.1 The Purpose of Citations

Scientific citations play a crucial role in maintaining the network of relationships among articles within a research field by linking together works whose methods and results are in some way mutally relevant. Customarily, 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. A researcher may then use the presence of citations to locate articles she needs to know about when entering a new field or to read in order to keep track of progress in a field where she is already well-established. But, determining the nature of the exact relationship between a citing and cited paper, whether a particular article is relevant and, if so, in what way, is often difficult to ascertain. To address this problem, the aim of citation analysis studies has been to categorize and, ultimately, automatically classify scientific citations. An automated citation classifier could be used, for example, in scientific indexing systems to provide additional information to help users navigating a digital library of scientific articles. In previous work, Garzone and Mercer [10] presented a system for citation classification that relied on a citation sentence's characteristic syntactic structure to assist in determining citation category. In this present work, we extend this idea to propose that fine-grained cue phrases within citation sentences may provide just such a stylistic basis for categorization.

1.2 Why Classify Citations? (and why this is difficult)

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. A citation connecting the source document and a citing document serves one of many functions. For example, one function is that the citing work gives some form of credit to the work reported in the source article. Another function is to criticize previous work. When using a citation index, a user normally has a more precise query in mind than "Find all articles citing a source article". Rather, the user may wish to know whether other experiments have used similar techniques to those used in the source article, or whether other works have reported conflicting experimental results. In order to use a citation index in this more sophisticated manner, the citation index must contain not only the citation-link information, but also must indicate the function of the citation in the citing article. The function of the citation must therefore be determined using information derived from local and global cues in the citing article.

The use of citation indexing of scientific articles was invented by Dr. Eugene Garfield in the 1950s as a result of studies on problems of medical information retrieval and indexing of biomedical literature. Dr. Garfield later founded the Institute for Scientific Information (ISI), whose Science Citation Index [8] is now one of the most popular citation indexes. Recently, with the advent of digital libraries, Web-based indexing systems have begun to appear (e.g., ISI's 'Web of Knowledge', CiteSeer [1]).

In all cases, 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. Citations link researchers and related articles together, and allow navigation through a space of mutually relevant documents which define a coherent academic discipline. As an example, the ISI 'Web of Knowledge' "[maintains and improves] the access and links between users of scholarly information and additional repositories of relevant research..., integrating journal, patent, proceedings, and life science literature with Web resources and other scholarly content." (from the ISI Web site). In addition, however, citation indexes allow the importance of an article to be assessed based on the frequency and locations of the citations. Citation statistics can thus play an important role in academic affairs, including promotion and tenure decisions and research grant awards. For all these reasons, scientific citations are a crucial component in the research and administrative life of the academic community.

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 *categorizing* 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.

Many citation classification schemes have been developed, with great variance in the number and nature of categories used. Garfield [7] was the first to define a classification scheme, while Finney [5] was the first to suggest that a citation classifier could be automated. Other classification schemes include those by Cole [2], Duncan, Anderson, and McAleese [3], Frost [6], Lipetz [14], Moravcsik and Murugesan [17], Peritz [19], Small [20], Spiegel-Rösing [21], and Weinstock [23]. Within this representative group of classification schemes, the number of categories ranges from four to 26. Examples of these categories include a *contrastive*, *supportive*, or *corrective* relationship between citing and cited works.

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[3] ([9], [10]) 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 to good performance on previously unseen (fair performance) and previously seen (good performance) articles. We now propose to extend this idea to develop a method for using more finely-grained cue phrases within citation sentences as a stylistic basis for categorization.

1.3 Background to the Research

Garzone and Mercer As Garzone and Mercer ([9], [10]) demonstrated, the problem of classifying citation contexts can be based on the recognition of certain *cue words* or

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

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

Teufel As a basis from which to develop our own approach to the citation problem, both the supporting work (i.e., Garzone and Mercer) and the opposing camp (e.g., Teufel) are useful references from which to start. In direct contrast to Garzone and Mercer, Teufel [22] questions whether fine-grained discourse cues do exist in citation contexts, and states that "many instances of citation context are linguistically unmarked." (p. 93). She goes on to add that while "overt cues" may be recognized if they are present, the problems of detecting these cues by automated means are formidable (p. 125):

- One could use simple, short, well-edited texts with standardized punctuation.
- One could use task-structured texts.
- One could posit an "evidence oracle".
- One could perform "deep" intention modelling and recognition.

Teufel thus articulates the dual challenges facing us: to demonstrate that fine-grained discourse cues can play a role in citation analysis, and that such cues may be detected by automated means.

While Teufel does represent a counterposition to Garzone and Mercer, which we take as our starting-point, nevertheless her work lays important foundations for ours in a number of ways. Most importantly, Teufel acknowledges the importance of a recognizable rhetorical structure in scientific articles, the so-called 'IMRaD' structure, for *Introduction*, *Method*, *Results*, and *Discussion*. In her own work, which is aimed at generating summaries of scientific articles, Teufel relies on rhetorical structure as a means of determining where to find specific types of information to construct her 'fixed-form' summaries. In addition, Teufel builds from this very global discourse structure a very detailed model of scientific argumentation that she proposes using as a basis for analyzing and summarizing the content of an article, including citation content. This model consists of 31 argumentative 'moves', which are typically one clause or sentence in length, and which build, step by step, the rhetorical structure of the scientific presentation. Examples of argumentative moves include motivating the need for the current research by pointing out a weakness in previous work (p. 84), or continuing a tradition from other research (p. 85). Of interest to us, these argumentative moves are often reminiscent of the kinds of categories used in citation classification schemes, and, significantly, Teufel observes that an important assumption is that "the argumentative status of a certain move is visible on the surface by linguistic cues." (p. 84)

At this point, Teufel diverges from us in her development of a method for analyzing the structure of articles based on a detailed discourse model and fine-grained linguistic cues. She does nonetheless give many instances of argumentative moves that may be signalled in citation contexts by specific cues, which are underlined in the following examples (p. 92):

(2) CUG (Categorial Unification Grammar; Uszkoreit (1986)) <u>is advantageous</u>, compared to other phrase structure grammars, for parallel architecture, because we can regard categories as functional types and we can represent grammar rules locally. (An example of the argumentative move *showing other solution is advantageous.*)

(3) We present a different method that takes <u>as starting point</u> the back-off scheme of Katz (1987). (An example of argumentative move *stating other solution provides basis for own solution.*)

Teufel acknowledges her concern with the "potentially high level of subjectivity" (p. 92) inherent in judging the nature of citations, a task made more difficult by the fine granularity of her model of argumentation and the absence, she claims, of reliable means of mapping from citations to the author's reason for including the citation: "[articles] often contain large segments, particularly in the central parts, which describe research in a fairly neutral [i.e., unmarked] way." (p. 93) As a consequence, Teufel reduces her model to a computationally tractable, but very broad-based set of seven categories, and confines the citation categories to only two types: the cited work either provides a basis for the citing work or contrasts with it.

2 The Role of Discourse Structure in Citation Analysis

2.1 Our Approach: Using Detailed 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., [4], [11], [18]), 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 [13] and Marcu [16], 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 4, 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.

(4) 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 [13] 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 [16], 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 also adopt:

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:

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

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

(6) 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:

(7) 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:

(8) 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:

(9) 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) [15] provides the current most popular set of rhetorical relations for use in Computational Linguistics. Marcu [16] 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 (p. 142); 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 The First Step: Determining the Frequency of Cue Phrases in Citations

The underlying premise of studies on the role of cue phrases in discourse structure (e.g., [12], [13], [16]) 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[4] articles. Our current task is to test our hypothesis that fine-grained discourse cues do exist in citation contexts. The details of the first stages of this analysis are presented in the next sections. Our analysis confirms that cue phrases do occur in citation contexts with about the same frequency as their occurrence in the complete text.

Description of the Analysis We are using a dataset of 24 scholarly science articles. All of these articles are written in the IMRaD style. (Four articles have merged the Results and Discussion sections into a single section.) We are using the list of cue phrases from [16] in our analysis. Our belief that this list is adequate for this initial analysis results from the fact that it is an extension of the one from [13], which was derived from academic text.

We analyze the use of cue phrases in three components of the article. The first component is the *full text body*. The full text body starts with the Introduction section and finishes with the Discussion section (or the merged Results and Discussion section). We also subdivide the full text body into its four (or three) sections, Introduction, Methods, Results, and Discussion. The second component is the *citation sentence* which is any sentence in the full text body that contains at least one citation. The third component is the *citation window*. Each citation window corresponds to a citation sentence together with the preceding and following sentences. When citation windows overlap, the citation windows are merged. Hence, a citation window may contain more than one citation sentence. When the citation sentence is the first or last sentence of one of the IMRaD sections, the missing preceding or following sentence is not included.

For each of these components, the number of words is counted and the number of times each cue phrase is used is tabulated, giving the frequency of cue-phrase usage. Also tabulated are the number of citation sentences in the full text body and in each IMRaD section.

Results of the Analysis The results of our analysis are given in the following three tables. We provide the details for each paper rather than a summary, since it is instructive at this point to see how the papers vary in the various statistics.

Table 1 shows the frequency of citation sentences in the full text body and the fraction of the citations occurring in each of the sections. Articles r1200, r3557, r432, and

[4] We are currently working with one scientific genre, biochemistry.

r4446 have their Results and Discussion sections merged. The table shows that between one-tenth and one-fifth of the sentences in these articles are citation sentences, with an average of 0.14. That citation sentences comprise between one-tenth and one-fifth of the sentences in a scientific article helps to demonstrate our earlier statement about the importance of making connections to extra-textual information. We contend that writers of scientific text use the same linguistic techniques to maintain cohesion between the textual and extra-textual material as they do to make their paper cohesive. The importance of these techniques, which we mentioned earlier, and the simple fact that their linguistic signals occur as frequently in citation sentences as in the rest of the text, which we discuss below, lends positve weight to our hypothesis, contra Teufel, that fine-grained discourse cues do exist in citation contexts and that they are relatively simple to find automatically.

The remaining columns tabulate the fraction of citation sentences in each section. Citations are well-represented in each of the IMRaD sections suggesting that a purpose exists for relating each aspect of a scientific article to extra-textual material. Further analysis is required to catalogue these relationships and how they are signalled.

Tables 2 and 3[5] corroborate our hypothesis that cue phrases do exist in citation contexts. In addition, the frequency of their occurrence suggests that cue phrases do play a significant role in citations: we note that the usage of cue phrases in citation sentences and citation windows is about the same as the usage in the full text body.

Another interesting feature that may be seen in these tables is that cue-phrase usage in the Methods section is lower (one insignificant higher value), and sometimes significantly lower, than cue-phrase usage in the full text body. One of our hypotheses is that the rhetoric of science will be part of our understanding of text cohesion in this type of writing. The Methods section is highly stylized, often being a sequence of steps. Further analysis may reveal that this rhetorical style obviates the use of cue phrases in certain situations.

In addition to our global frequency analysis that we have given above, it is important to analyze the frequency of individual cue phrases. In Table 4 we show just a few instances from the 60 most frequently occurring cue phrases to point out some interesting patterns.

The cue phrase *previously* is three times more frequent in citation sentences than in the full text body and twice as frequent as in citation windows. This may indicate a strong tendency to indicate temporal coherence. The cue phrase *not* is used 50% more frequently in textcitation windows than in citations. Does this show that citation windows set up negative contexts? Similarly, *however* appears almost 50% more frequently in textcitation windows than in citations. Similar 'opposites' for *although*, *following*, and *in order to* seem to be present in the data.

4 Conclusions and Future Work

Our primary concern was to find evidence that fine-grained discourse cues exist in significant number in citation contexts. Our analysis of 24 scholarly science articles indi-

[5] The cue phrase *and* is often used as a coordinate conjunction. We removed this word from the list of cue phrases to see if the results differed. If anything, the result was stronger.

X

Table 1. Citation sentence occurrence.

			Frequency of cue sentence usage								
	Num of	Num of		Cit in		Cit in		Cit in		Cit in	
Article	Sent	Cit	Freq	Intro	Freq	Meth	Freq	Res	Freq	Disc	Freq
r1182	240	36	0.15	4	0.11	14	0.39	7	0.19	11	0.31
r1200	358	42	0.12	8	0.19	9	0.21	25		0.60	
r1265	305	53	0.17	22	0.42	2	0.04	7	0.13	11	0.21
r1802	233	36	0.15	14	0.39	9	0.25	5	0.14	8	0.22
r1950	401	62	0.15	10	0.16	14	0.23	22	0.35	16	0.26
r1974	226	28	0.12	11	0.39	4	0.14	6	0.21	7	0.25
r1997	358	52	0.15	12	0.23	16	0.31	9	0.17	15	0.29
r2079	222	32	0.14	8	0.25	10	0.31	5	0.16	9	0.28
r2603	198	26	0.13	12	0.46	4	0.15	4	0.15	6	0.23
r263	436	70	0.16	17	0.24	2	0.03	16	0.23	35	0.50
r315	275	33	0.12	9	0.27	11	0.33	8	0.24	5	0.15
r3343	251	40	0.16	14	0.35	10	0.25	6	0.15	10	0.25
r3557	202	23	0.11	8	0.35	6	0.26	9		0.39	
r3712	349	47	0.13	14	0.30	12	0.26	11	0.23	10	0.21
r3819	420	43	0.10	13	0.30	9	0.21	14	0.33	7	0.16
r432	288	40	0.14	10	0.25	7	0.17	23		0.57	
r4446	365	56	0.15	11	0.20	16	0.29	29		0.52	
r5007	402	58	0.14	17	0.29	11	0.19	15	0.26	15	0.26
r513	266	52	0.20	13	0.25	8	0.15	16	0.31	15	0.29
r5948	276	57	0.21	9	0.16	11	0.19	14	0.25	23	0.40
r5969	445	64	0.14	21	0.33	6	0.09	11	0.17	26	0.41
r6200	256	25	0.10	5	0.20	8	0.32	3	0.12	9	0.36
r7228	301	32	0.11	9	0.28	7	0.22	0	0.00	16	0.50
r7903	218	35	0.16	9	0.26	10	0.29	6	0.17	10	0.29

Table 2. Cue phrase occurrence.

Frequency of cue phrase usage in various contexts							
Article	Full Body	Introduction	Methods	Results	Discussion	Citation	Cit Win
r1182	0.119	0.119	0.093	0.119	0.110	0.115	0.103
r1200	0.096	0.093	0.078	0.098		0.089	0.093
r1265	0.102	0.096	0.078	0.110	0.099	0.095	0.095
r1802	0.096	0.072	0.072	0.116	0.122	0.107	0.089
r1950	0.100	0.112	0.089	0.092	0.103	0.100	0.095
r1974	0.108	0.088	0.075	0.104	0.128	0.107	0.096
r1997	0.102	0.103	0.103	0.095	0.103	0.114	0.098
r2079	0.112	0.077	0.093	0.105	0.127	0.123	0.103
r2603	0.093	0.106	0.064	0.080	0.103	0.111	0.089
r263	0.123	0.132	0.096	0.109	0.133	0.115	0.123
r315	0.107	0.106	0.078	0.096	0.131	0.105	0.091
r3343	0.103	0.111	0.090	0.086	0.109	0.106	0.098
r3557	0.115	0.105	0.081	0.117		0.125	0.109
r3712	0.092	0.076	0.090	0.085	0.094	0.088	0.087
r3819	0.114	0.112	0.097	0.106	0.118	0.114	0.107
r432	0.101	0.098	0.075	0.101		0.102	0.107
r4446	0.109	0.090	0.102	0.106		0.086	0.094
r5007	0.104	0.106	0.093	0.095	0.115	0.102	0.107
r513	0.099	0.091	0.094	0.089	0.099	0.089	0.093
r5948	0.121	0.114	0.103	0.112	0.127	0.117	0.117
r5969	0.102	0.098	0.069	0.093	0.109	0.092	0.099
r6200	0.100	0.105	0.074	0.097	0.100	0.101	0.094
r7228	0.104	0.082	0.093	0.092	0.111	0.106	0.103
r7903	0.103	0.095	0.099	0.082	0.108	0.095	0.094

Table 3. Cue phrase occurrence (with “and” removed from the list of cue phrases).

Frequency of cue phrase usage in various contexts (“and” removed)							
Article	Full Body	Introduction	Methods	Results	Discussion	Citation	Cit Win
r1182	0.093	0.094	0.062	0.095	0.087	0.094	0.079
r1200	0.063	0.059	0.044	0.069		0.061	0.063
r1265	0.069	0.060	0.054	0.069	0.084	0.065	0.068
r1802	0.068	0.049	0.044	0.096	0.098	0.082	0.064
r1950	0.072	0.084	0.055	0.069	0.086	0.080	0.070
r1974	0.080	0.067	0.038	0.078	0.106	0.088	0.076
r1997	0.066	0.080	0.062	0.058	0.073	0.081	0.066
r2079	0.077	0.050	0.067	0.077	0.085	0.088	0.072
r2603	0.071	0.079	0.043	0.065	0.081	0.080	0.065
r263	0.094	0.107	0.057	0.081	0.107	0.091	0.101
r315	0.078	0.080	0.049	0.069	0.108	0.084	0.066
r3343	0.080	0.079	0.061	0.071	0.090	0.075	0.073
r3557	0.072	0.081	0.043	0.076		0.094	0.075
r3712	0.066	0.051	0.062	0.060	0.077	0.061	0.063
r3819	0.089	0.085	0.068	0.084	0.098	0.086	0.082
r432	0.070	0.056	0.049	0.074		0.066	0.076
r4446	0.079	0.062	0.078	0.075		0.065	0.067
r5007	0.076	0.073	0.066	0.070	0.090	0.072	0.080
r513	0.074	0.069	0.061	0.065	0.081	0.069	0.073
r5948	0.098	0.101	0.069	0.087	0.115	0.099	0.099
r5969	0.072	0.070	0.034	0.070	0.081	0.065	0.071
r6200	0.071	0.075	0.042	0.077	0.080	0.063	0.063
r7228	0.076	0.042	0.059	0.071	0.092	0.078	0.075
r7903	0.072	0.063	0.066	0.055	0.086	0.063	0.068

Table 4. Frequencies of example cue phrases.

Citation sentences		Citation windows		Full text body	
100	0.0316 previously	110	0.0170 previously	124	0.0102 previously
78	0.0246 not	199	0.0308 not	404	0.0333 not
28	0.0088 although	49	0.0076 although	70	0.0058 although
22	0.0069 however	63	0.0097 however	116	0.0096 however
11	0.0035 following	30	0.0046 following	78	0.0064 following
6	0.0019 in order to	16	0.0025 in order to	36	0.0030 in order to

cates that these cues do exist in citation contexts, and that their frequency is comparable to that in the full text. Secondarily, we are very interested in whether these cues are automatically detectable. Many of these discourse cues appear as cue phrases that have been previously catalogued in both academic and general texts. The detection of these cue phrases has been shown to be straigthforward. What may be of equal importance are discourse cues that are not members of the current list of cue phrases: we envisage 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. There may be other (discourse) cue phrases characteristic of scientific writing and citationese: Knott ([13]) has suggested two categories—propositional anaphora and sentential complements that introduce an intentional stance—that appear to be used quite frequently in citation style. In addition, there may be other types of citationese cue phrases entirely: 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 openings).

Of course, the main goal of this study of discourse relations is to use the linguistic cues 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 coherencerhetorical relations signalled by cue phrases to extra-textual coherence relations linking citing and cited papers?

References

1. Bollacker, B., Lawrence, S., and Giles, C.L.: A system for automatic personalized tracking of scientific literature on the Web. In *Digital Libraries 99—The Fourth ACM Conference on Digital Libraries*. ACM Press, New York (1999) 105–113
2. Cole, S.: The growth of scientific knowledge: Theories of deviance as a case study. In *The Idea of Social Structure: Papers in Honor of Robert K. Merton*, Harcourt, Brace Jovanovich, New York (1975) 175–220
3. Duncan, E.B., Anderson,F.D., and McAleese, R.: Qualified citation indexing: its relevance to educational technology. In *Information retrieval in educational technology: Proceedings of the first symposium on information retrieval in educational technology*, University of Aberdeen (1981) 70–79
4. Fahnestock, J.: *Rhetorical figures in science.* Oxford University Press (1999)
5. Finney, B.: The reference characteristics of scientific texts. Master's thesis, The City University of London (1979)
6. Frost, C.: The use of citations in literary research: a preliminary classification of citation functions. *Library Quarterly*, Volume 49, (1979) 399–414
7. Garfield, E.: Can citation indexing be automated? In M.E. Stevens et al., editors, *Statistical Association Methods for Mechanical Documentation (NBS Misc. Pub. 269)*. National Bureau of Standards, Washington, DC (1965)
8. Garfield, E.: Information, power, and the *Science Citation Index*. In *Essays of an Information Scientist*, Volume 1, 1962–1973, Institute for Scientific Information.
9. Garzone, M.: *Automated classification of citations using linguistic semantic grammars*. M.Sc. Thesis, The University of Western Ontario (1996)

10. Garzone, M., and Mercer, R.E.: Towards an automated citation classifier. In *AI'2000, Proceedings of the 13th Biennial Conference of the CSCSI/SCEIO*, Lecture Notes in Artificial Intelligence, v. 1822, H. J. Hamilton (ed.), Springer-Verlag, (2000) 337–346
11. Gross, A.G.: *The rhetoric of science.* Harvard University Press (1996)
12. Halliday, M.A.K., and Hasan, Ruqaiya.: *Cohesion in English.* Longman Group Limited (1976)
13. Knott, A.: *A data-driven methodology for motivating a set of coherence relations*. Ph.D. thesis, University of Edinburgh (1996)
14. Lipetz, B.A.: Problems of citation analysis: Critical review. *American Documentation*, Volume 16 (1965) 381–390
15. Mann, W.C., and Thompson, S.A.: Rhetorical structure theory: Toward a functional theory of text organization. *Text*, Volume 8(3) (1988)
16. Marcu, D.: *The rhetorical parsing, summarization, and generation of natural language texts*. Ph.D. thesis, University of Toronto (1997)
17. Moravscik, M.J., and Murugesan, P.: Some results on the function and quality of citations. *Social Studies of Science*, Volume 5 (1975) 86–92
18. Myers, G.: *Writing biology.* University of Wisconsin Press (1991)
19. Peritz, B.C.: A classification of citation roles for the social sciences and related fields. *Scientometrics*, Volume 5 (1983) 303–312
20. Small, H.: Cited documents as concept symbols. *Social Studies of Science*, Volume 8(3) (1978) 327–340
21. Spiegel-Rösing, I.: Science studies: Bibliometric and content analysis. *Social Studies of Science*, Volume 7 (1977) 97–113
22. Teufel, S.: *Argumentative zoning: Information extraction from scientific articles*. Ph.D. thesis, University of Edinburgh (1999)
23. Weinstock, M.: Citation indexes. In *Encyclopaedia of Library and Information Science*, Volume 5, Marcel Dekkar, New York (1971) 16–40