Chapter 19

Using Hedges to Classify Citations in Scientific Articles

Chrysanne Di Marco and Frederick W. Kroon
Dept. Of Computer Science, University of Waterloo,
Waterloo, Ontario, Canada
Email*: cdimarco@uwaterloo.ca*

Robert E. Mercer
Dept. Of Computer Science, The University of Western Ontario,
London, Ontario, Canada
Email*: mercer@csd.uwo.ca*

Abstract

Citations in scientific writing fulfil an important role in creating relationships among mutually relevant articles within a research field. These inter-article relationships reinforce the argumentation structure intrinsic to all scientific writing. Therefore, determining the nature of the exact relationship between a citing and cited paper requires an understanding of the rhetorical relations within the argumentative context in which a citation is placed. To determine these relations automatically, we have suggested that various stylistic and rhetorical cues will be significant. One such cue that we are studying is the use of hedging to modify the affect of a scientific claim. We provide evidence that hedging occurs more frequently in citation contexts than in the text as a whole. With this information we conjecture that hedging is a significant aspect of the rhetorical structure of citation contexts and that the pragmatics of hedges may help in determining the rhetorical purpose of citations. A citation indexing tool for biomedical literature analysis is introduced.

Keywords: automatic citation analysis, hedges, rhetoric of science, science writing.

1. Scientific Writing, the Need for Affect, and Its Role in Citation Analysis

Since the inception of the formal scientific article in the seventeenth century, the process of scientific discovery has been inextricably linked with the actions of writing and publishing the results of research. Rhetoricians of science have gradually moved from a purely descriptive characterization of the science genre to full-fledged field studies detailing the evolution of the

scientific article. During the first generation of rhetoricians of science, (e.g., Myers, 1991, Gross, 1996, Fahnestock, 1999), the persuasive nature of the scientific article, how it contributes to making and justifying a knowledge claim, was recognized as the defining property of scientific writing. Style (lexical and syntactic choice), presentation (organization of the text and display of the data), and argumentation structure were noted as the rhetorical means by which authors build a convincing case for their results. Recently, second-generation rhetoricians of science (e.g., Hyland, 1998, Gross et al., 2002) have begun to methodically analyze large corpora of scientific texts with the purpose of cataloguing specific stylistic and rhetorical features that are used to create the pragmatic effects that contribute to the author's knowledge claim. One particular type of pragmatic effect, *hedging*, is especially common in scientific writing and can be realized through a wide variety of linguistic choices.

We believe that pragmatic attitudes such as hedging (Hyland, 1998), politeness (Myers, 1989), and persuasion play an essential role in building the argumentative structure of the scientific article, and in conveying the nuances that help to support the author's knowledge claims. Moreover, we believe that these pragmatic effects work together with both global discourse structure—e.g., the traditional Introduction, Methods, Results, and Discussion (IMRaD) design of scientific discourse—and local text structure, including lexical choice, syntactic arrangement, citation placement and other aspects of scientific presentation, to create the overall rhetorical effect of a research article. In particular, we are studying the pragmatic function of citations in providing a textual means of relating articles in the space of documents which defines a research community. Studies in citation analysis indicate that the author's intent in including a citation at a particular point in the text reflects the pragmatic purpose of the citation, whether, for example, it indicates supporting or contrasting work to the topic under discussion. Our basic hypothesis is that the specific pragmatic function of citations may be determined through the analysis of fine-grained linguistic cues in the surrounding text.

We are presently studying the analysis of hedging cues in scientific writing as a means of classifying the purpose of citations in scientific texts. Hedging analysis seems well-suited as a means of approaching this problem: hedging in scientific writing is both pervasive and often readily detectable by surface textual features, while hedging cues have been well-studied (e.g., Hyland, 1998) in terms of their pragmatic function.

We have started to apply our citation classification methodology in the biomedical field. We believe that the usefulness of automated citation classification in literature indexing can be found in both the larger context of managing entire databases of scientific articles and for specific information-extraction problems such as mining the literature for protein-protein interactions.

2. Hedging in Scientific Writing

Hyland (1998) elaborates on "hedging", the term introduced by Lakoff (1972) to describe "words whose job it is to make things more or less fuzzy.": "[Hedging] has subsequently been applied to the linguistic devices used to qualify a speaker's confidence in the truth of a proposition, the kind of caveats like *I think*, *perhaps*, *might*, and *maybe* which we routinely add to our statements to avoid commitment to categorical assertions. Hedges therefore express tentativeness and possibility in communication, and their appropriate use in scientific discourse is critical (p1)".

The following examples illustrate some of the ways in which hedging may be used to deliberately convey an attitude of uncertainty or qualification. In the first example, the use of the verb *suggested* hints at the author's hesitancy to declare the absolute certainty of the claim:

(1) The functional significance of this modulation is suggested by the reported inhibition of MeSo-induced differentiation in mouse erythroleukemia cells constitutively expressing c-myb.

In the second example, the syntactic structure of the sentence, a fronted adverbial clause, emphasizes the effect of qualification through the rhetorical cue *Although*. The subsequent phrase, *a certain degree*, is a lexical modifier that also serves to limit the scope of the result:

(2) Although many neuroblastoma cell lines show a certain degree of heterogeneity in terms of neurotransmitter expression and differentiative potential, each cell has a prevalent behavior in response to differentiation inducers.

Hedging may be used in different rhetorical contexts within a scientific article to convey persuasive effect and enhance the knowledge claims of the author. For example, hedging may be realized through various linguistic cues in the Introduction, Results section, a controversial Discussion section, or generally throughout the research paper.

Within the Introduction to a scientific article, the use of hedging may serve both to establish the results within a wider research context and highlight the significance of this new work. In the extract below, the authors repeatedly use the key phrase *is/are consistent with* to first establish the reliability of their results, and then turn to more-hesitant cues (*provide circumstantial evidence*, *may be responsible*, *Regardless of the validity of this specific proposal*) to support, yet not overreach, their assertions. Nevertheless, the authors still manage to get their claims across through a number of subtle but signficant cues: *not appear to*, *we reasoned*, *would*.

(3) Transgenic Arabidopdis seedlings over expressing phytochrome B exhibit enhanced sensitivity to Rc but wild-type responsiveness to FRc (Wagner *et al*, 1991; McCormac *et al*, 1993). This result is consistent with the behaviour of endogenous phytochrome B deduced from the *hy* 3 mutant studies...By contrast, transgenic Arabidopdis over expressing phytochrome A exhibits enhanced sensitivity to FRc (Whitelam *et al*, 1992; McCormac *et al*, 1993). Together these results are consistent with the possibility, although do not prove, that the capacity to mediate the FR-HIR may be an intrinsic property of phytochrome A.

Accumulated biochemical and physiological data also provide circumstantial evidence that phytochrome A may be responsible for the FR-HIR...[the data] are consistent with the possibility that this photolabile phytochrome pool may be responsible for the FR-HIR.

Regardless of the validity of this specific proposal, however, because phytochrome B does not appear to be involved in the FR-HIR, we reasoned that mutants defective in the activity of the phytochrome mediating this response would retain phytochrome B, and, therefore, retain responsiveness to Rc...

The Results section of a scientific paper, whether implicit or set off as a formal structure, tends to be lengthy and subdivided according to topic (Hyland, 1998, p193). The topics present the paper's findings, while associated hedges may be used to enhance the persuasive effects of the authors' interpretations of the findings and the resulting claims.

In the following example, the authors appear to be hedging certainty, putting forth their claim, but tempering the persuasive effect. They have chosen a modal verb, *would*, rather than a strong positive verb, such as *indicates*, so that the effect of the claim is restrained. Then, the following sentence seems to signal the possibility of a strong contrast by the explicit discourse marker,

However, and use of a negative phrase, *cannot be ruled out*. Overall, the rhetorical effect is one of hesitance and tentativeness on the author's part.

(4) The faint 21-kD band observed in the PBM lane (Figure 2) would reflect the transient passage of this protein across the PBM from the plant cell cytoplasm to the bacteroids. However, the opposite is also possible, and it cannot be ruled out that the 21-kD polypeptides seen in the bacteroid lane and in the soluble proteins lane are totally different proteins with the same apparent molecular weight.

Hedging may be used not only in enhancing or mitigating the persuasive effects of an author's specific knowledge claims, but in setting up a strong 'protective' position from which to defend a highly controversial position. Hyland (1998, p196) describes a text in which the writer has proposed a radical explanation for a process that is a core issue in her research area. As he analyzes the text, he points out how the writer goes even further, in making serious challenges to current theories. Not only is the writer concerned about supporting her own scientific claim, Hyland observes, but with protecting her position in her research community: "In making this proposal, the writer implicitly attributes serious inadequacies in current theories in their interpretations of critical data. She therefore runs the very real risk of having the claim rejected by a community of peers who, she perceives, have a great deal invested in the existing view and who are likely to defend it without giving serious consideration to her work" (p. 196).

How then does this writer manage to simultaneously put forth her own claim, challenge established theory, and protect her position in the community? Not surprisingly, the paper is thick with hedges: modal verbs and adverbs, epistemic lexical verbs, indefinite quantifiers, and admissions of limiting conditions, all contriving to "[create] a rhetorical and interpersonal context which seeks to pre-empt the reader's rejection" (Hyland, 1998, p196).

As these examples illustrate, hedging effects are commonly used throughout scientific articles, while the ways in which hedging may be realized are both varied and easy to recognize. These characteristics suggested to us that the detection of hedging effects might be used as the basis for locating linguistic cues in scientific texts that might then help to determine the intended communicative effect of citations placed in the surrounding text.

3. Classifying Citations in Scientific Writing

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, the aim of formal citation analysis has been to categorize and, ultimately, automatically classify scientific citations.

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 (5) 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).

(5) 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).

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. Citation statistics play an important role in academic affairs, including promotion and tenure decisions and research grant awards. Scientific citations are thus 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. 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. In previous work, Garzone and Mercer (Garzone, 1996, Garzone and Mercer, 2000) presented a system for citation classification that relied on characteristic syntactic structure to determine citation category. We are now extending this idea to develop a method for using fine-grained rhetorical cues within citation sentences to provide such a stylistic basis for categorization (Mercer and Di Marco, 2003, Di Marco and Mercer, 2003, Mercer et al., 2004).

3.1 Related Work in Citation Classification

The usefulness of citation categorization for other applications is directly related to the comprehensiveness (breadth and granularity) of the citation classification scheme. Garzone and Mercer (Garzone, 1996, Garzone and Mercer, 2000) proposed a citation classification scheme with 35 categories. This scheme is more comprehensive than the union of all of the previous schemes: it has a finer granularity than the often-used scheme of Garfield (1965) and Weinstock (1971) and the one which previously had the most categories, (Duncan et al., 1981), and it includes the full breadth of the other schemes (Cole, 1975, Finney, 1979, Frost, 1979, Lipetz, 1965, Moravscik and Murugesan, 1975, Peritz, 1983, Small, 1982, and Spiegel-Rösing, 1977). The Garzone and Mercer scheme and its relationship to the previous ones is discussed in detail in Garzone (1996).

We list a few of the citation categories (slashes indicate separate categories):

- Citing work disputes/corrects/questions some aspect of cited work.
- Citing work confirms/illustrates some aspect of cited work.
- Use of materials, equipment, or tools/methods, procedures, and design/theoretical equation/definition/numerical data.

We have a prototype citation classification system that takes journal articles (currently only biochemistry and physics) as input and maps each citation into one of the 35 citation categories. The prototype system relies on a large number of cue words (for example, discourse cues, nouns, and verbs which are closely related to the science and its methodology), some simple syntactic relationships, and knowledge about the IMRaD structure.

In direct contrast to Garzone and Mercer, which we take as our own starting-point, Teufel (1999) questions whether fine-grained discourse cues do exist in citation contexts, and states that "many instances of citation context are linguistically unmarked." (p93). She adds that while "overt cues" may be recognized if they are present, the problems of detecting these cues by automated means are formidable (p125). 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 represents a counterposition to our approach, her work does complement ours in a number of ways. Teufel's research has a different goal to ours – it is aimed

at generating summaries of scientific articles – but she does acknowledge the importance of a recognizable discourse structure in scientific articles, the IMRaD structure, and she also relies on local rhetorical structure to help determine where to find specific types of information to construct her 'fixed-form' summaries. However, Teufel voices her concern about the "potentially high level of subjectivity" (p92) 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. As a consequence, Teufel confines her classification of citation categories to only two clearly distinguishable types: the cited work either provides a basis for the citing work or contrasts with it.

Nanba and Okumura (1999) and Nanba et al. (2000) also present work in automated citation classification that is complementary to ours: their aim is to automatically generate review articles in a specific subject domain using citation types as the basis for the classification of papers. Like Teufel, they rely on two primary citation categories (works that provide a supporting basis for the citing paper, works that have a contrasting or 'negative' relationship), but also add a third 'others' category to indicate some form of unspecified relationship exists between the citing and cited papers. Collections of 'cue phrases' (including discourse markers, lexical usage, specific phrases), are used to classify citations into the different categories but these cues are heuristically motivated rather than theoretically based. In contrast, the types of cues we are using to detect the purpose of a citation are based in discourse analysis (Mercer and Di Marco, 2003) and the rhetoric of science (Mercer et al., 2004).

We can thus summarize the differences between our approach to citation categorization and that of Teufel and Nanba et al. as follows:

- Our aim is a literature indexing tool using the rhetoric of science.
- We use a fine-grained citation categorization scheme with a greater number and variety of categories.
- We rely on cue phrases derived from formal linguistic theories as the basis for the detection and classification of citations.

4. Determining the Importance of Hedges in Citation Contexts

The surface features through which hedging is realized in scientific texts have been copiously catalogued, in particular by Hyland. Using several corpora, both scientific and general academic, Hyland (1998) carried out a detailed analysis of hedging at several levels of linguistic description, including surface-level cataloguing of hedges and pragmatic analysis of their functions (pp98–99). The results of the study yielded a detailed catalogue of hedging cues including a large number of modal auxiliaries, epistemic lexical verbs (most commonly, *suggest*, *indicate*, *predict*), epistemic adjectives, adverbs, and nouns (representing half the major grammatical classes expressing hedging), as well as a variety of non-lexical, discourse-based hedges.

We believe that hedging cues may provide a prime source of fine-grained discourse cues that can be used to determine the intent of citations in the surrounding text. Hedging cues seem ideally suited for this purpose because the various types of hedging in scientific discourse have been extensively studied and catalogued by rhetoricians of science, (Hyland, 1998), in particular, and because the surface cues that give rise to hedging are readily recognizable by linguistic analysis, e.g., modal auxiliaries, specific lexical choice, and the use of discourse markers.

In our initial study (Mercer and Di Marco, 2003), we analyzed the frequency of discourse cues in a set of scholarly scientific articles. We 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. We noted in this study that citations appeared to occur quite often in sentences marked by hedging cues. For example, sentence (1) above contains the hedging verb *suggested*, and a citation about earlier work by other authors. We may assume that the hedge and the citation are linked in some way: hesitancy in the current work may be offset by the support of earlier related research.

In sentence (2) above, the lexical and syntactic cues (*Although*, *a certain degree*) express qualification of the claim, but now the accompanying use of several citations serves to bolster the authoritative nature of the underlying argument. (Indeed, two of the citations refer to papers published more than five years earlier, and the third reference is 17 years old.)

Frame Sentence	<p> To test this idea further, we also analyzed a construct where the third Val residue in the V18 segment was changed to Pro.
Citation Sentence	We have previously shown that the introduction of a Pro residue in corresponding positions in a L23V transmembrane segment leads to a reduction in the MGD value of about 2.5 residues, presumably as a result of a break in the poly-Leu -helix caused by the Pro residue [<citation/>14].
Frame Sentence	Indeed, the initial drop in the glycosylation profile for the V18(P3) construct was ~2 residues, Fig. 4B, while the shift in the location of the second drop was only ~1 residue.
Normal Sentence	This is consistent with the possibility that V18 molecules with MGD ~ 15.5 residues indeed have already formed a transmembrane-helix at the time of glycosylation, whereas the remaining ones have not.

Figure 1. A paragraph (starts with <p>) containing all sentence types. There are two hedge cues (underlined) in this example, one in the citation frame, and one outside the citation window.

We have followed up on our hypothesis that hedging cues tend to occur in citation contexts with a frequency analysis of hedging cues in citation contexts in a 985 biology journal article subset from the BioMed Central corpus, and obtained statistically significant results indicating that hedging is indeed used more frequently in citation contexts than the whole text (Mercer et al., 2004). Given the presumption that writers make stylistic and rhetorical choices purposefully, we propose this as further evidence that hedging cues are an important aspect of the rhetorical structure of citation contexts and the pragmatic functions of hedges may help to determine the purpose of citations.

Each sentence in the corpus was identified as one or more of the following (see Figure 1):

- A citation sentence, if the sentence contains one or more citations.
- A citation frame sentence, if the sentence contains no citation and is immediately adjacent to a citation sentence that is within the same paragraph.
- A normal sentence, if it is neither a citation nor a citation frame sentence.
- A hedge sentence, if the sentence contains one or more hedging cues.

Several tallies were computed. We kept track of each citation sentence and frame, noting whether each contained a hedging cue. In addition, each citation window, which comprises both the citation sentence and the citation frame, was noted as either containing or lacking a hedging cue. Finally, we tallied the total number of sentences that contain a hedging cue, the total number of sentences that contain a citation, and the total number of sentences that fall into a citation frame.

It was often the case that citation windows overlapped in the text. This is especially evident in the citation-rich background section. When this occurred, care was taken to avoid double-counting hedging cues. When a hedging cue occurred in the intersecting region of two citation windows, the cue was counted as belonging to only one of the two windows. If it was in the citation sentence of one of the two windows, it was counted as belonging to the citation sentence in which it fell. If it fell in the intersection of two citation frames, it was counted as belonging to the citation that had no other hedge within its window. If neither window contained any other hedging cues, it was arbitrarily treated as belonging to the first of the two windows.

Table 1 shows the counts and Table 2 shows the frequencies of citation sentences, frame sentences, and hedge sentences. Any given sentence may belong to only one of the citation/frame categories. Since citation windows may overlap, it is sometimes the case that a citation sentence may also be part of the frame of another window. In this case, the sentence is counted only once, as a citation sentence, and not as a citation-frame sentence. Note that in Table 2, the frequencies do not add to 1, since there are sentences that neither occur in a citation window nor contain hedging cues. Data about these sentences has not been listed in Table 2.

	Total	Citation		Hedge Sentences		
Section	Sentences	Sentences	Frames	Verb	Non-verb	Total
background	22321	10172	6037	2891	2785	5278
methods	36632	5922	5585	2132	1480	3468
results+disc	87382	16576	16405	13602	12040	23198
conclusions	5145	587	647	1049	760	1635

Table 1. Number of sentences, by sentence type.

	Citation		Hedge Sentences		
Section	Sentences	Frames	Verb	Non-verb	Total
background	0.46	0.27	0.13	0.12	0.24
methods	0.16	0.15	0.06	0.04	0.09
results+disc	0.19	0.19	0.16	0.14	0.27
conclusions	0.11	0.13	0.20	0.15	0.32

Table 2. Proportion of total sentences, by sentence type.

	Verb Cues			Non-verb Cues			All Cues		
Section	Cite	Frame	All	Cite	Frame	All	Cite	Frame	All
background	0.15	0.11	0.13	0.13	0.13	0.12	0.25	0.22	0.24
methods	0.09	0.06	0.06	0.05	0.04	0.04	0.14	0.10	0.09
results+disc	0.22	0.16	0.16	0.15	0.14	0.14	0.32	0.27	0.27
conclusions	0.29	0.22	0.20	0.18	0.19	0.15	0.42	0.36	0.32

Table 3. Proportion of sentences containing hedging cues, by type of sentence and hedging cue category.

Hedge sentences are further subdivided into verb and non-verb categories depending on whether the hedging cue is a verb or a non-verb. Note that a sentence may belong to both of these categories. The reason for this is that the sentence may contain two cues, one from each category. In all cases, a sentence containing more than one hedging cue is counted only once as a hedge sentence (reported in the 'Total' column). This single-counting of sentences containing multiple cues explains why the number of hedge sentences does not total to the number of hedging cues.

Table 3 shows the proportions of the various types of sentences that contain hedging cues, broken down by hedging-cue category. For all but two combinations, citation sentences are more likely to contain hedging cues than would be expected from the overall frequency of hedge sentences at a significance level of 0.01. The two combinations for which there are no significant differences are non-verb hedging cues in the background and conclusion sections. It is interesting to note that there are, however, significantly (at a significance level of 0.01) more non-verb cues than expected in citation frames in the conclusion section.

With the exception of the above combination (non-verb cues in the conclusion section), citation frame sentences seem to contain approximately the same proportion of hedging cues as the overall text. However, this being said, there is little indication that they contain fewer cues than expected. The one major exception to this trend is that citation frame sentences in the background section appear less likely to contain verbal hedging cues than would be expected. It is not clear whether this is due to an actual lack of cues, or is simply an artifact of the fact that since the background section is so citation rich, there are relatively few citation frames counted (since a sentence is never counted as both a citation sentence and a citation frame sentence).

	n		Verb Cues		Non-verb Cues		All Cues	
Section	citation	frame	citation	frame	citation	frame	citation	frame
background	10172	6037	32.66	22.19	0.97	0.93	15.69	5.65
methods	5922	5585	118.75	0.94	13.53	0.03	113.82	1.33
results+disc	16576	16405	451.48	0.58	20.53	2.01	288.36	4.19
conclusions	587	647	24.50	1.17	5.57	9.92	26.86	6.16

Table 4. Chi2(1,n) values for observed versus expected proportion of citation sentences and frames containing hedging cues. Chi2 (crit) is 9.14 after Bonferroni correction.

	Windows		Sentences		Frames	
Section	#	%	#	%	#	%
background	3361	0.33	2575	0.25	2679	0.26
methods	1089	0.18	801	0.14	545	0.09
results+disc	7257	0.44	5366	0.32	4660	0.28
conclusions	338	0.58	245	0.42	221	0.38

Table 5. Number and proportion of citation windows containing a hedging cue, by section and location of hedging cue.

	Windows		Sentences		Frames	
Section	#	%	#	%	#	%
background	1967	0.19	1511	0.15	1479	0.15
methods	726	0.12	541	0.09	369	0.06
results+disc	4858	0.29	3572	0.22	2881	0.17
conclusions	227	0.39	168	0.29	139	0.24

Table 6. Number and proportion of citation windows containing a verbal hedging cue, by section and location of hedging cue.

	Windows		Sentences		Frames	
Section	#	%	#	%	#	%
background	1862	0.18	1302	0.13	1486	0.15
methods	432	0.07	295	0.05	198	0.03
results+disc	3751	0.23	2484	0.15	2353	0.14
conclusions	186	0.32	107	0.18	111	0.19

Table 7. Number and proportion of citation windows containing a non-verb hedging cue, by section and location of hedging cue.

The chi2(1,n) values for observed versus expected proportion of citation sentences and frame sentences containing hedging cues are summarized in Table 4. The chi2(1,n) values were computed by comparing the actual versus expected frequencies of hedging cues in each sentence type. The expected frequencies are obtained simply from the overall frequency of each sentence type. Thus, if hedging cues were distributed randomly, and 24% of sentences overall had hedging cues, one would expect that approximately 24% of citation sentences would contain cues, assuming there is no relationship between hedging and citations. In order to correct for multiple tests, Bonferroni correction (Miller, 1981) was applied.

Tables 5, 6, and 7 summarize the occurrence of hedging cues in citation windows. Table 8 shows the proportion of hedge sentences that either contain a citation, or fall within a citation frame. Note that this is not the same thing as the proportion of *hedging cues* that fall within a citation sentence or frame. If more than one hedging cue falls within a single sentence, the sentence is counted as a single hedge sentence.

	Verb Cues			Non-verb Cues			All Cues		
Section	Cite	Frame	None	Cite	Frame	None	Cite	Frame	None
background	0.52	0.23	0.25	0.47	0.28	0.25	0.49	0.26	0.26
methods	0.25	0.16	0.59	0.20	0.15	0.65	0.23	0.16	0.61
results+disc	0.26	0.19	0.55	0.21	0.19	0.60	0.23	0.19	0.58
conclusions	0.16	0.14	0.70	0.14	0.16	0.70	0.15	0.14	0.71

Table 8. Proportion of hedge sentences that contain citations or are part of a citation frame, by section and hedging cue category.

Table 8 suggests (last 3-column column) that the proportion of hedge sentences containing citations or being part of citation frame is at least as great as what would be expected just by the distribution of citation sentences and citation windows. Table 3 indicates that in most cases the proportion of hedge sentences in the citation windows is greater than what would be expected by the distribution of hedge sentences. Taken together, these conditional probabilities support the conjecture that hedging cues and citation contexts correlate strongly. Rather than occurring by chance, writers purposefully use these cues. With this knowledge, the strong correlation would indicate that the hedging cues are being used in synergy with the citation contexts. Hyland has catalogued a variety of pragmatic uses of hedging cues, so it is reasonable to speculate that these uses map over to the rhetorical structure that is found in citation contexts.

5. A Citation Indexing Tool for Biomedical Literature Analysis

We are presently developing a biomedical literature indexing tool to automate the classification of citations using the rhetoric of science through the following tasks:

- Adapting existing computational linguistic tools (e.g., online lexicons, part-of-speech taggers, discourse marker analyzers) for the detection of hedging cues and other cue phrases within citation contexts.
- Building test corpora of citation sentences from biomedical and scientific articles.
- Developing methods and tools for automatically classifying the pragmatic functions of hedging cues and other cue phrases in the citation corpora.

Our goal in studying the effects of hedging in scientific writing is to identify linguistic cues that may be used as a means of determining the pragmatic function of citations. Ultimately, we can expect to be able to associate hedging cues and other pragmatic cues with rhetorical relations as determiners of citation function.

Indexing tools, such as CiteSeer (Bollacker et al., 1999), play an important role in the scientific endeavour by providing researchers with a means of navigating through the network of scholarly scientific papers using the connections provided by citations. Citations relate articles within a research field by linking together works whose methods and results are in some way mutually 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. Another 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, with the explosion in the amount of scientific literature, a means of providing more information in order to give more intelligent control to the navigation process is warranted. A user normally wants to navigate more purposefully 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 navigate 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. But, the author's purpose for including a citation is not apparent in the citation per se. Determining the nature of the exact relationship between a citing and cited paper, often requires some level of understanding the text that the citation is embedded in.

The goal of our citation indexing tool project is the design and implementation of an indexing tool for scholarly biomedical literature which uses the text surrounding the citation to provide information about the binary relation between the two papers connected by a citation. In particular, we are interested in how the scientific method structures the way in which ideas, results, theories, etc. are presented in scientific writing and how the style of presentation indicates the purpose of citations, that is, what the relationship is between the cited and citing papers. Our interest in the connections among scientific literature (our focus), ontologies, and databases is that the content and structure of each of these three repositories of scientific knowledge has its foundations in the method of science.

A *citation index* enables 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 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 (Garfield, 1973) 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' (http://www.isinet.com), CiteSeer (Bollacker et al., 1999)).

In the biomedical field, we believe that the usefulness of automated citation classification in literature indexing can be found in both the larger context of managing entire databases of scientific articles or for specific information-extraction problems. On the larger scale, database curators need accurate and efficient methods for building new collections by retrieving articles on the same topic from huge general databases. Simple systems (e.g., Andrade and Valencia, 1988, Marcotte et al., 2001) consider only keyword frequencies in measuring article similarity. More-sophisticated systems, such as the Neighbors utility (Wilbur and Coffee, 1994), may be able to locate articles that appear to be related in *some* way (e.g., finding related Medline abstracts for a set of protein names (Blaschke et al., 1999), but the lack of specific information about the nature and validity of the relationship between articles may still make the resulting collection a less-than-ideal resource for subsequent analysis. Citation classification to indicate the nature of the relationships between articles in a database would make the task of building collections of related articles both easier and more accurate. And, the existence of additional knowledge about the nature of the linkages between articles would greatly enhance navigation among a space of documents to retrieve meaningful information about the related content.

A specific problem in information extraction that may benefit from the use of citation categorization involves mining the literature for protein-protein interactions (e.g., Blaschke et al., 1999, Marcotte et al., 2001, Thomas et al., 2000). Currently, even the most-sophisticated systems are not yet capable of dealing with all the difficult problems of resolving ambiguities and detecting hidden knowledge. For example, Blaschke et al.'s system (Blaschke et al., 1999) is able to handle fairly complex problems in detecting protein-protein interactions, including constructing the network of protein interactions in cell-cycle control, but important implicit knowledge is not recognized. In the case of cell-cycle analysis for *Drosophila*, their system is able to determine that relationships exist between **Cak**, **Cdk7**, **CycH**, and **Cdk2**: **Cak** inhibits/phosphorylates **Cdk7**, **Cak** activates/phosphorylates **Cdk2**, **Cdk7** phosphorylates **Cdk2**, **CycH** phosphorylates **Cak** and **CycH** phosphorylates **Cdk2**. However, the system is not able to detect that **Cak** is actually a complex formed by **Cdk7** and **CycH**, and that the **Cak** complex regulates **Cdk2**. While the earlier literature describes inter-relationships among these proteins, the recognition of the generalization in their structure, i.e., that these proteins are part of a complex, is contained only in more-recent articles: "There is an element of generalization implicit in later publications, embodying previous, more dispersed findings. A clear improvement here would be the generation of associated weights for texts according to their level of generality" (Blaschke et al., 1999). Citation categorization could provide just these kind of 'ancestral' relationships between articles—whether an article is foundational in the field or builds directly on closely related work—and, if automated, could be used in forming collections of articles for study that are labelled with explicit semantic and rhetorical links to one another. Such collections of semantically linked articles might then be used as 'thematic' document clusters (cf. (Wilbur, 2002)) to elicit much more meaningful information from documents known to be closely related.

An added benefit of having citation categories available in text corpora used for studies such as extracting protein-protein interactions is that more, and more-meaningful, information may be obtained. In a potential application, Blaschke et al. (1999) noted that they were able to discover many more protein-protein interactions when including in the corpus those articles found to be related by the Neighbors facility (Wilbur and Coffee, 1994) (285 versus only 28 when relevant protein names alone were used in building the corpus). Lastly, very difficult problems in scientific and biomedical information extraction that involve aspects of deep-linguistic meaning may be

resolved through the availability of citation categorization in curated texts: synonym detection, for example, may be enhanced if different names for the same entity occur in articles that can be recognized as being closely related in the scientific research process.

5.1 Our Guiding Principles

The automated labelling of citations with a specific citation function requires an analysis of the linguistic features in the text surrounding the citation, coupled with a knowledge of the author's pragmatic intent in placing the citation at that point in the text. The author's purpose for including citations in a research article reflects the fact that researchers wish to communicate their results to their scientific community in such a way that their results, or *knowledge claims*, become accepted as part of the body of scientific knowledge. This persuasive nature of the scientific research article, how it contributes to making and justifying a knowledge claim, is recognized as the defining property of scientific writing by rhetoricians of science, (e.g., Gross, 1996, Gross et al., 2002, Hyland, 1998, Myers, 1991). Style (lexical and syntactic choice), presentation (organization of the text and display of the data), and argumentation structure are noted as the rhetorical means by which authors build a convincing case for their results.

Our approach to automated citation classification is based on the detection of fine-grained linguistics cues in scientific articles that help to communicate these rhetorical stances and thereby map to the pragmatic purpose of citations. As part of our overall research methodology, our goal is to map the various types of pragmatic cues in scientific articles to rhetorical meaning. Our previous work has described the importance of *discourse cues* in enhancing inter-article cohesion signalled by citation usage (Mercer and Di Marco, 2003, Di Marco and Mercer, 2003). We have also been investigating another class of pragmatic cues, *hedging cues*, (Mercer et al., 2004), that are deeply involved in creating the pragmatic effects that contribute to the author's knowledge claim by linking together a mutually supportive network of researchers within a scientific community.

5.2 Our Design Methodology

The indexing tool that we are designing is an enhanced citation index. The feature that we are adding to a standard citation index is the function of each citation, that is, given an agreed-upon set of citation functions, we want our tool to be able to automatically categorize a citation into one of these functional categories. To accomplish this automatic categorization we are using a decision tree—currently, we are building the decision tree by hand, but in future we intend to investigate machine learning techniques to induce a tree. Our aim is to have a working indexing tool whenever we add more knowledge to the categorization process. This goal appears very feasible given our design methodology choice of using a decision tree: adding more knowledge only refines the decision-making procedure of the previously working version.

Two factors influence the development of the tree as follows:

- the granularity of the categories determines the number of leaves in the decision tree
- the number of features used to categorize determines the potential depth of the tree.

We are using Garzone and Mercer's 35-category scheme (Garzone, 1996, Garzone and Mercer, 2000) in the citation classifier, but a finer or coarser granularity is obviously permitted. Concerning the features on which the decision tree makes its decisions, we have started with a

simple, yet fully automatic prototype (Garzone, 1996) which takes journal articles as input and classifies every citation found therein into at least one of the 35 categories. Its decision tree is very shallow, using only sets of cue-words and polarity switching words (not, however, etc.), some simple knowledge about the IMRaD structure of the article together with some simple syntactic structure of the citation-containing sentence. In addition to having a design which allows for easy incorporation of more-sophisticated knowledge, it also gives flexibility to the tool: categories can be easily coalesced to give users a tool that can be tailored to a variety of uses.

Although we anticipate some small changes to the number of categories due to category refinement, the major modifications to the decision tree will be driven by a more-sophisticated set of features associated with each citation. When investigating a finer granularity of the IMRaD structure, we came to realize that the structure of scientific writing at all levels of granularity was founded on *rhetoric*, which involves both argumentation structure and stylistic choices of words and syntax. This was the motivation for choosing the rhetoric of science as our guiding principle.

We rely on the notion that rhetorical information is realized in linguistic 'cues' in the text, some of which, although not all, are evident in surface features (cf. Hyland, 1998) on surface hedging cues in scientific writing. Since we anticipate that many such cues will map to the same rhetorical features that give evidence of the text's argumentative and pragmatic meaning, and that the interaction of these cues will likely influence the text's overall rhetorical effect, the formal *rhetorical relation* (cf. (Mann and Thompson, 1988)) appears to be the appropriate feature for the basis of the decision tree. So, our long-term goal is to map between the textual cues and rhetorical relations. Having noted that many of the cue words in the prototype are discourse cues, and with two recent important works linking discourse cues and rhetorical relations (Knott, 1996, Marcu, 1997), we began our investigation of this mapping with these cues. We have some early results that show that discourse cues are used extensively with citations and that some cues appear much more frequently in the citation context than in the full text (Mercer and Di Marco, 2003). Another textual device is the hedging cue, which we are currently investigating (Mercer et al., 2004).

Although our current efforts focus on cue words which are connected to organizational effects (discourse cues), and writer intent (hedging cues), we are also interested in other types of cues that are associated more closely to the purpose and method of science. For example, the scientific method is, more or less, to establish a link to previous work, set up an experiment to test an hypothesis, perform the experiment, make observations, then finally compile and discuss the importance of the results of the experiment. Scientific writing reflects this scientific method and its purpose: one may find evidence even at the coarsest granularity of the IMRaD structure in scientific articles. At a finer granularity, we have many targetted words to convey the notions of procedure, observation, reporting, supporting, explaining, refining, contradicting, etc. More specifically, science categorizes into taxonomies or creates polarities. Scientific writing then tends to compare and contrast or refine. Not surprisingly, the morphology of scientific terminology exhibits comparison and contrasting features, for example, *exo-* and *endo-*. Science needs to measure, so scientific writing contains measurement cues by referring to scales (0–100), or using comparatives (larger, brighter, etc.). Experiments are described as a sequence of steps, so this is an implicit method cue.

Finally, as for our prototype system, we will continue to evaluate the classification accuracy of the citation-indexing tool by a combination of statistical testing and validation by human experts. In addition, we would like to assess the tool's utility in real-world applications such as database curation for studies in biomedical literature analysis. We have suggested earlier that there may be

many uses of this tool, so a significant aspect of the value of our tool will be its ability to enhance other research projects.

6. Conclusions and Future Work

In this paper we have motivated our hypothesis that hedging cues should and can be exploited in the process of determining the nature of citation function, and our approach to developing a literature indexing tool that computes the functions of citations. The function of a citation is determined by analyzing the rhetorical intent of the text that surrounds it. This analysis is founded on the guiding principle that the scientific method is reflected in scientific writing. The purposeful nature of citation function is a feature of scientific writing which can be exploited in a variety of ways. We anticipate more-informative citation indexes as well as more-intelligent database curation. Additionally, sophisticated information extraction may be enhanced when better selection of the dataset is enabled. For example, synonym detection in a corpus of papers may be made more tractable when the corpus is comprised of related papers derived from navigating a space of linked citations. Our early investigations have determined that linguistic cues and citations are related in important ways. Our future work will be to map these linguistic cues to rhetorical relations and other pragmatic functions so that this information can then be used to determine the purpose of citations

7. Acknowledgements

Our research has been financially supported by the Natural Sciences and Engineering Research Council of Canada and by the Universities of Western Ontario and Waterloo.

8. Bibliography

Andrade, M. A., and Valencia, A. (1988) Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. *Bioinformatics*, 14(7), 600-607.

Blaschke, C., Andrade, M. A., Ouzounis, C., and Valencia, A. (1999) Automatic extraction of biological information from scientific text: Protein-protein interactions. In *Int. Conf. on Intelligent Systems for Molecular Biology (ISMB)*, 60-67.

Bollacker, B., Lawrence, S., and Giles, C. L. (1999) A system for automatic personalized tracking of scientific literature on the Web. In *The Fourth ACM Conf. on Digital Libraries,* 105-113.

Cole, S. (1975) 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, New York, 175-220.

Di Marco, C., and Mercer, R. E. (2003) Toward a catalogue of citation-related rhetorical cues in scientific texts. In *Proc. of the Pacific Assoc. for Comp. Ling. Conf. (PACLING)*, 63-72.

Duncan, E. B., Anderson, F. D., and McAleese, R. (1981) Qualified citation indexing: its relevance to educational technology. In *Information retrieval in educational technology,* 70-79.

Fahnestock, J. (1999) *Rhetorical figures in science.* Oxford University Press.

Finney, B. (1979) The reference characteristics of scientific texts. Master's thesis, The City University of London.

Frost, C. (1979) The use of citations in literary research: a preliminary classification of citation functions. *Library Quarterly*, 49, 399-414.

Garfield, E. (1965) 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.

Garfield, E. (1973) Information, power, and the *Science Citation Index*. In *Essays of an Information Scientist*, 1, 1962–1973, Institute for Scientific Information.

Garzone, M. (1996) *Automated classification of citations using linguistic semantic grammars.*, M.Sc. Thesis, The University of Western Ontario.

Garzone, M., and Mercer, R. E. (2000) Towards an automated citation classifier. In *Proc. of the Conf. of the Canadian Society for the Computational Studies of Intelligence (CSCSI)*, 337-346.

Gross, A. G. (1996) *The rhetoric of science.* Harvard University Press.

Gross, A. G., Harmon, J. E., and Reidy, M. (2002) *Communicating science: The scientific article from the 17th century to the present.* Oxford University Press.

Hyland, K. (1998) *Hedging in scientific research articles.* John Benjamins Publishing Company.

Knott, A. (1996) *A data-driven methodology for motivating a set of coherence relations.* Ph.D. thesis, University of Edinburgh.

Lakoff, R. (1972) The pragmatics of modality. In P. Peranteau, J. Levi, and G. Phares, editors, *Papers from the Eighth Regional Meeting*, Chicago Linguistics Society, 229-246.

Lipetz, B. A. (1965) Problems of citation analysis: Critical review. *Am. Doc.*, 16, 381-390.

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

Marcotte, E. M., Xenarios, I., and Eisenberg, D. (2001) Mining literature for protein-protein interactions. *Bioinformatics*, 17(4), 359-363.

Marcu, D. (1997) *The rhetorical parsing, summarization, and generation of natural language texts.* Ph.D. thesis, University of Toronto.

Mercer, R. E., and Di Marco, C. (2003) The importance of fine-grained cue phrases in scientific citations. In *Proc. of the Conf. of the Can. Soc. for the Comp. Studies of Int. (CSCSI),* 550-556.

Mercer, R. E., Di Marco, C., and Kroon, F. W. (2004) The frequency of hedging cues in citation contexts in scientific writing. In *Proc. of the Conf. of the Canadian Society for the Computational Studies of Intelligence (CSCSI),* 75-88.

Miller, R. G. (1981) *Simultaneous statistical inference*, Springer Verlag.

Moravscik, M. J., and Murugesan, P. (1975) Some results on the function and quality of citations. *Social Studies of Science*, 5, 86–92.

Myers, G. (1989) The pragmatics of politeness in scientific articles. *Appl. Linguistics,* 10(1), 1-35.

Myers, G. (1991) *Writing biology.* University of Wisconsin Press.

Nanba, H. and Okumura, M. (1999) Towards multi-paper summarization using reference information. In *Proc. of the 16th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 926-931.

Nanba, H., Kando, N., and Okumura, M. (2000) Classification of research papers using citation links and citation types: Towards automatic review article generation. In *Proc. of the American Society for Information Science (ASIS)*, 117-134.

Peritz, B. C. (1983) A classification of citation roles for the social sciences and related fields. *Scientometrics*, 5, 303-312.

Small, H. (1982) Citation content analysis. *Progress in Communication Sciences*, 3, 287-310.

Spiegel-Rösing, I. (1977) Science studies: Bibliometric and content analysis. *Social Studies of Science*, 7, 97-113.

Teufel, S. (1999) *Argumentative zoning: Information extraction from scientific articles.* Ph.D. thesis, University of Edinburgh.

Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carroll, M. (2000) Automatic extraction of protein interactions from scientific abstracts. In *Proc. of the 5th Pacific Symp. on Biocomputing (PSB)*, 538-549.

Weinstock, M. (1971) Citation indexes. In *Encycl. of Library and Information Science*, 5, 16-40.

Wilbur, W. J. (2002) A thematic analysis of the AIDS literature. In *Proc. of the 7th Pacific Symp. on Biocomputing (PSB)*, 386-397.

Wilbur, W. J., and Coffee., L. (1994) The effectiveness of document neighboring in search enhancement. *Information Processing Management*, 30, 253-266.