The frequency of hedging cues in citation contexts in scientific writing

Robert E. Mercer[1], Chrysanne Di Marco[2], and Frederick W. Kroon[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, fwkroon@uwaterloo.ca

Abstract. Citations in scientific writing fulfill an important role in creating relationships among mutually relevant articles within a research field. These inter-article relationships reinforce the argumentation structure that is 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 in scientific writing, we have suggested that stylistic and rhetorical cues will be significant. One type of cue that we have studied is the discourse cue, which provides cohesion among textual components. Another form of rhetorical cue involves hedging to modify the affect of a scientific claim. Hedging in scientific writing has been extensively studied by Hyland, including cataloging the pragmatic functions of the various types of cues. In this paper we show that the hedging cues proposed by Hyland occur more frequently in citation contexts than in the text as a whole. With this information we conjecture that hedging cues are an important aspect of the rhetorical relations found in citation contexts and that the pragmatics of hedges may help in determining the purpose of citations.

1 Introduction

1.1 Why We Are Studying Hedging

Citations in scientific writing are used to connect mutually relevant articles within a research field. A citation index, which is a compilation of connections between citing and cited articles, provides a means to navigate the collection of scientific articles. Typically, a navigational foray into this collection is overwhelmed by the sheer numbers of related articles. However, users of citation indexes usually have a more focussed purpose than merely to find related articles. Very often, the index search could be narrowed if the writer's purpose for generating the citation were to label the citation connection.

To label citations with a citation function drawn from a list of possible functions requires an analysis of the text surrounding the citation coupled with the knowledge that scientific researchers want to communicate their results and want

to argue that these results become part of the body of scientific knowledge. This latter aspect has been extensively studied by rhetoricians of science, e.g. [5], [6], [12].

These rhetorical stances are communicated with a variety of stylistic effects. Fundamental to our work is the claim that these stylistic effects may be realized through surface cues in text thereby enabling the reader to interpret the rhetorical nature of the text appropriately. With this in mind, our overall task is to map stylistic cues to rhetorical relations. Our previous work has described the importance of *discourse cues* in enhancing inter-article cohesion signalled by citation usage [11]. We have also begun to compile a catalogue of fine-grained discourse cues that exist in citation contexts [1]. In the following we investigate another class of pragmatic cues signalled by surface means—called 'hedging'—that can be exploited to uncover the function of citations in scientific writing.

The hypothesis that we test is that hedging cues play an important role in the rhetoric that is intrinsic to citation usage. We have investigated this hypothesis by doing a frequency analysis of hedging cues in citation contexts in a corpus of 985 biology articles. We have obtained statistically significant results indicating that hedging is used more frequently in citation contexts than the text as a whole. Given the presumption that writers make stylistic and rhetorical choices purposefully, we propose that we have further evidence that connections between fine-grained linguistic cues and rhetorical relations exist and that these may be catalogued. Further, the rhetorical relation can then suggest a citation function.

1.2 Why Hedging is Used in Scientific Writing

The 'job' of a scientific researcher is usually thought to be focussed primarily on the discovery of new factual knowledge. However, a large and critical component of scientific discovery concerns the acceptance of new results by the research community and their ultimate integration into the community's archival knowledge. The scientific process is therefore not only about *doing* science but about *persuading* others of the validity of results so that these may be judged worthy of publication. Along with the validation and publication of results, the reputation of the individual researcher becomes enhanced, with concomitant effects on the person's standing in her community, chances of receiving tenure or promotion, and likely success in subsequent grant reviews.

A variety of rhetorical strategies may be used in scientific writing to create both a sense of social involvement in the scientific community and a persuasive influence in promoting the reader's acceptance of the claims being made. The rhetorical means through which an author achieves this persuasive purpose may take various forms: *hedging*, to weaken the assertiveness of new claims, so they may be more readily judged acceptable by reviewers; *citations*, to indicate a network of mutually supportive or contrasting works; and *politeness* to build social closeness with other researchers [8] (p. 64). Of these, the rhetorical strategies that will concern us here are the uses of hedging and citations, and the extent to which these two strategies are linked.

Hedging strategies may be described as "...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 [8] (p. 1)". The use of hedging in scientific writing is actually part of a larger pragmatic purpose on the part of the author: she is simultaneously putting forth claims that must be seen as worthy of publication or as a basis for funding, while at the same time she must be careful to present her work as acceptable to her social community of academic peers and as constituting a continuation of established knowledge (even though she may be to some extent challenging or replacing previous work). Hyland [8] (p. 196) 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). To address these conflicting pragmatic purposes, 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" [8] (p. 196).

In attempting to persuade readers of a research article that the results—*knowledge claims*—contained therein constitute a valuable addition to the discipline, the author is in effect engaging in an intricate 'dialogue' with her audience. Hedging can thus be viewed as a type of modality that allows

> a form of participation by the speaker in the speech event. Through modality, the speaker associates with the thesis an indication of its status and validity in his own judgement; he intrudes, and takes up a position. [7] (p. 335), quoted in [8] (p. 47)

We may say therefore that hedging as a rhetorical technique in building up a scientific argument is intrinsic to scientific writing. Further, the pragmatic functions of hedging, conveying persuasive effect to enhance new knowledge claims and aiding the writer in building a social context with the reader, would seem to indicate that effectively managing the use of hedging is essential to the scientific process. As Hyland [8] states in his review of hedging strategies in scientific writing:

> Hedging is critical in scientific discourse because it helps gain communal acceptance for knowledge. Scientific 'truth' is as much a social as an intellectual category, and the distinction writers make between

> their subject matter and how they want readers to understand their relationship to it is crucial to such a highly self-conscious form of discourse. Not only does it influence the effectiveness and credibility of argumentation, but helps define what it means to write science.... [8] (p. 38)

We take as our guiding principle the thesis that these larger pragmatic functions of hedging indicate that other rhetorical strategies in scientific writing—for example, the use of citations—may be found to work together with hedges in creating persuasive effects and influencing interpersonal judgements.

1.3 Background to the Research

A *citation* may be formally defined as a portion of a sentence in a citing document which references another document or a set of other documents collectively. A *citation sentence* is any sentence in the full text body that contains at least one citation. A *citation window* corresponds to a citation sentence together with the preceding and following sentences, if they occur in the same paragraph.

Garzone and Mercer Garzone and Mercer [4] motivated the current citation categorization project. This foundational work demonstrated a classifier system based on a correspondence between certain cue words, specific word usages, and characteristic structural patterns in citing sentences and the citation functions performed by the 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.

(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).

Di Marco and Mercer Di Marco and Mercer [1] developed the first stages of a method for citation classification guided by the hypothesis that the fine-grained rhetorical structure of a scientific article can help tremendously in this task. This hypothesis is based on the following two arguments:

- The well-established body of work in rhetorical theory may be used in analyzing the global structure of scientific discourse, e.g., [3], [5], [6], [12].
- More-recent studies have demonstrated the role of fine-grained discourse cues [9] [10] in the rhetorical analysis of general text.

We are thus developing an approach to citation classification in which the recognition of such subtle linguistic cues, 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. As a key part of our methodology, we intend to show that a direct

mapping can be determined from the pragmatic functions of these linguistic cues to the rhetorical purpose of the citations in the context within which they are both used.

In our preliminary study [11], we analyzed the frequency of the cue phrases from [10] 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. In subsequent work [1], we analyzed the same dataset of articles to begin to catalogue the fine-grained discourse cues that exist in citation contexts. This study confirmed that authors do indeed have a rich set of linguistic and non-linguistic methods to establish discourse cues in citation contexts.

We found that several types of syntactic stylistic usage provide rhetorical cues that may serve to indicate the nature of the citation. For example, the use of syntactic symmetry or parallelism can act as a cue for the enumeration of one or more citations. Repetition of words and phrases may also be considered a form of lexical 'parallelism' that occurs along with citations. Other forms of rhetorical cueing rely on various kinds of lexical stylistic features within a citation context: lexical morphology (e.g., contrasting concepts, such as *intra*cellular and *extra*cellular, that set up for a 'contrasting-work' citation); specific lexical choice (e.g., negative or 'extreme' words to describe results that seem to call for citations to supporting works); scientific procedural terms; and 'reporting' verbs (e.g., to make reference to the historical record of the author's own or other works). For this latter category of reporting cues, we observed the use of hedging verbs used along with a reporting style in citation contexts (e.g., *it has been previously suggested...* that led us to investigate the relationship between hedging and citation occurrences in more detail.

Teufel Teufel [14] represents a direct contrast to the approach taken by Garzone and Mercer and, by extension, our own approach. Teufel's work is concerned with the automated generation of summaries of scientific articles using the rhetorical structure of the document to find specific types of information to fill slots in a 'fixed-form' summary template. Teufel proposes a detailed model of scientific argumentation that may be used as the 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.

Teufel diverges from us in questioning whether citations are in any way linguistically marked, and, in particular, whether fine-grained discourse cues even occur in citation contexts. Even if such "overt cues" do exist, she notes, the task of detection through automated means would be formidable, requiring either deep-linguistic analysis or use of only simple, short, well-edited texts. Teufel thus articulates the dual challenges facing us: to demonstrate that fine-grained linguistic cues can in fact play a role in citation analysis and that such cues can be detected by automated means.

Although Teufel's approach runs counter to ours in judging whether citation contexts may be classified on the basis of subtle linguistic markers, she does nonetheless give many instances of argumentative moves that may be signalled in citation contexts by specific cues. Of interest to us, Teufel's set of argumentative moves is 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) However, Teufel voices her concern with the "potentially high level of subjectivity" (p. 92) inherent in judging the nature of citations. As a consequence, she relies on only two clearly distinguishable citation categories in developing her ultimate argumentative model : the cited work either provides a basis for the citing work or contrasts with it. In our approach, we hope to maintain both a very broad and varied set of citation categories, while at the same time developing methods for reliable citation classification based on the automated detection of subtle but pragmatically well-defined rhetorical cues.

2 The Frequency of Hedging Cues in Citation Contexts

The argumentative structure found in scientific writing is supported by a variety of rhetorical techniques, including hedging. Underlying our interest in studying citations in scientific writing is the supposition that citation use is a rhetorical strategy. The work reported here begins to investigate the rhetorical links between citation usage and rhetorical cues. The hypothesis that we test is that hedges play an important role in the rhetoric that is intrinsic to citation usage. We investigate this hypothesis by doing a frequency analysis of hedging in citation contexts in a corpus of scholarly biology articles.

We set out to compare the frequencies of hedging cues occurring in citation sentences and the sentences immediately surrounding the citation sentences to the frequency of hedging cues in the text as a whole. If these frequencies show a statistically significant difference, these differences may provide evidence for our hypothesis. A list of the hedging cues can be found in Appendix A. Writers make purposeful stylistic and rhetorical choices, therefore, we propose that frequency differences would be supporting evidence that hedging plays a role in the rhetoric that is intrinsic to citation usage. Demonstrating our hypothesis, in addition to providing further evidence to support Hyland's work, would give us reason to study the rhetorical relationship between hedging and citations at a much more detailed level.

3 Methodology

The corpus that was analyzed is a 985-document subset of the BioMed Central corpus[3] which is composed of biomedical journal articles. Only journal articles deemed to have a strong biological focus (as opposed to a medical focus) were

[3] http://www.biomedcentral.com/

included in the test corpus. All articles were required to have the following sections: background, methods, results, discussion, and conclusion. This structure is very similar to the IMRaD[4] structure used by Teufel [14]. Since many of the articles merged the results and discussion sections, the two were treated as a single, aggregate section.

The corpus was first preprocessed to remove any extraneous XML tags. Most of these tags are either located in the front matter, or are for the purposes of formatting. Only information about paragraph boundaries, citation locations, and section boundaries was kept. The following example illustrates a typical body sentence from the corpus. Only the <p> tag, which denotes the start of a paragraph, and the <abbr> tag, which denotes a citation were needed.

(2) <p>Previous studies have examined the nucleotide length distribution of the 5' UTRs, 3' UTRs, intergenic regions and space between RBSs and start sites of transcription in the genome of <it>E. coli </it><abbrgrp><abbr bid="B5">5</abbr></abbrgrp>.

A major advantage of the BioMed Central corpus is that citations are explicitly marked through the use of the <abbr> tag. As such it was unnecessary to cope with the various possible citation styles. Furthermore, since we are only interested in the presence or absence of hedging cues in a sentence, no syntactic processing was required. Finally, no morphological processing was performed. Since the number of hedging cues is small, the list of hedging cues taken from Hyland's catalogue was simply expanded to include all inflectional variants. We did not search for all derivational variants, since there is some evidence that not all such variants are used for hedging purposes [8].

The corpus was then split into sentences using a maximum-entropy sentence boundary detector, called MXTERMINATOR[5], described in [13]. The model was trained on a manually segmented portion of the corpus, consisting of approximately 3000 sentences. Unfortunately, MXTERMINATOR did not perform well on the biological corpus, and significant manual post-processing was required to correct segmentation errors. It is not clear if this is a result of the small training-set size, or the increased complexity of biological articles as compared to the Wall Street Journal articles on which MXTERMINATOR was originally evaluated.

The manual post-processing consisted of searching the output text from MXTERMINATOR for indications that MXTERMINATOR likely made an error, then manually correcting the mistake. There were several such indicators, such as single-word sentences, capital letters finishing a sentence (problematic since terms like "E. coli" were often split across sentences), and citations located at the beginning of a sentence (since these should be associated with the previous sentence, not the subsequent one), and so on.

Once segmentation was complete, each sentence in the corpus was identified as one or more of the following:

[4] Introduction, Methods, Results, and Discussion
[5] http://www.cis.upenn.edu/~adwait/statnlp.html

- 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 hedge sentence, if the sentence contains one or more hedging cues.

Citations are only counted if they are citations to published work. Hence, citations of unpublished data, personal communications and so on are not counted as citations for our purposes. Such citations to unpublished sources are not marked up in the BioMed corpus, so no additional processing was needed to exclude them.

On occasion a citation may occur immediately following the period at the end of a sentence. In such instances, the citation is included as part of the immediately preceding sentence, rather than the following sentence. The following example illustrates such a case.

(3) Studies have shown highest apoptosis expression in lining epithelium at estrus in mouse <citation/> and rat. <citation/><citation/>

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.

4 Results and Discussion

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

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 $\sim$ 2 residues, Fig. 4B, while the shift in the location of the second drop was only $\sim$ 1 residue.
Normal Sentence	This is consistent with the possibility that V18 molecules with MGD $\sim$ 15.5 residues indeed have already formed a transmembrane-helix at the time of glycosylation, whereas the remaining ones have not.

Fig. 1. A paragraph containing all three sentence types. There are two hedge cues (underlined) in this example, one in the citation frame, and one outside the citation window.

window nor contain hedging cues. Data about these sentences has not been listed in Table 2.

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 do not add up to the total 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 ($p \leq .01$). 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 ($p \leq .01$) 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).

Table 1. Number of sentences, by sentence type.

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

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

	Citation		Hedge Sentences		
	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
res+disc	0.19	0.19	0.16	0.14	0.27
conclusions	0.11	0.13	0.20	0.15	0.32

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

	Verb Cues			Non-verb Cues			All Cues		
	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
res+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 4. $\chi^2(1,n)$ values for observed versus expected proportion of citation sentences and frames containing hedging cues. $\chi^2_{crit} = 9.14$ after Bonferroni correction.

	n		Verb Cues		Non-verb Cues		All Cues	
	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
res+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

The $\chi^2(1,n)$ values for observed versus expected proportion of citation sentences and frame sentences containing hedging cues are summarized in Table 4. $\chi^2(1,n)$ 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

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

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

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

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

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

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

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

	Verb Cues			Non-verb Cues			All Cues		
	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
res+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

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 χ^2 tests, Bonferroni correction 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.

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 Conclusions and Future Work

In creating inter-article relationships, citations in scientific writing reinforce the argumentation structure that is intrinsic to all scientific writing. To determine the relationship between a citing and cited paper, we are unravelling the argumentation structure by looking for fine-grained discourse and rhetorical cues that indicate the rhetorical relations that build the argumentation structure. One type of rhetorical cue involves hedging to modify the affect of a scientific claim. In this paper we show that the hedging cues proposed in Hyland's extensive study of the rhetorical use of hedging in scientific writing occur more frequently in citation contexts than in the text as a whole. We have also confirmed that hedging is distributed unevenly in the different sections of a scientific article. With this information we conjecture that hedging cues are an important aspect of the rhetorical relations found in citation contexts and that the pragmatics of hedges may help in determining the purpose of citations.

The pragmatic nature of hedges can be divided into several categories, most broadly, *content-oriented hedges* and *reader-oriented hedges*. Content-oriented hedges "hedge the correspondence between what the writer says about the world and what the world is thought to be like" [8] (p. 162). Reader-oriented hedges are concerned with the social relationship between writer and reader, and are used for such pragmatic purposes as reducing the potential risks of presenting controversial claims, showing courtesy or deference to peers, and demonstrating conformity to the research community's expectations [8] (pp. 177-178). In the following example, the use of a personal subject (*We*) functions to mitigate the criticism of other work through an "overt acceptance of personal responsibility" [8] (p. 181). In addition, the modal *might* further weakens the critical effect:

(4) We do not know the reason for the discrepancy between our results and those of Ngernprairitsiri [16, 23], but it might reflect genetic differences in the cultivars employed.

As this last example shows, the pragmatic purpose of hedges and citations occurring within the same passage can be closely linked. As we observed in our study, specific types of hedges were associated with citations that offset the implied uncertainty. One type of content-oriented hedge expresses deviation from established or idealized knowledge [8] (p. 164) and is often associated with a citation to foundational work. In the following example, the adverb *slightly* indicates that the procedure being used deviates from that of Buchner *et al.*:

(5) Drosophila heads were harvested and prepared according to a slightly modified protocol described in Buchner et al. [38].

We are now beginning to develop a mapping from the pragmatic functions of hedging cues to the purpose of citations used within the same context. Our ultimate purpose is to identify fine-grained linguistic cues that may be used as a means of determining the function of citations. Based on Hyland and others, we can expect to be able to associate hedging cues and other cue phrases with rhetorical relations as determiners of citation function.

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. Thanks also goes to Erin Harvey and Heather Thiessen in the Department of Statistics and Actuarial Science, University of Waterloo, for assistance with statistical design and analysis.

References

1. Di Marco, C. and R. E. Mercer: Toward a catalogue of citation-related rhetorical cues in scientific texts. In *Proceedings of the Pacific Association for Computational Linguistics Conference* (PACLING'03) (2003) 63–72
2. Di Marco, C. and R. E. Mercer: Hedging in scientific articles as a means of classifying citations. To appear in *AAAI Spring Symposium on Exploring Attitude and Affect in Text* (2004)
3. Fahnestock, J.: *Rhetorical figures in science.* Oxford University Press (1999)
4. Garzone, M. and R. E. Mercer: 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
5. Gross, A.G.: *The rhetoric of science.* Harvard University Press (1996)
6. Gross, A.G., J. E. Harmon, and M. Reidy: *Communicating science: The scientific article from the 17th century to the present.* Oxford University Press (2002)
7. Halliday, M. A. K.: Functional diversity in language as seen from a consideration of modality and mood in English. *Foundations of Language*, Volume 6, (1970) 322–361

8. Hyland, K.: *Hedging in scientific research articles.* John Benjamins Publishing Company (1998)
9. Knott, A.: *A data-driven methodology for motivating a set of coherence relations.* Ph.D. thesis, University of Edinburgh (1996)
10. Marcu, D.: *The rhetorical parsing, summarization, and generation of natural language texts.* Ph.D. thesis, University of Toronto (1997)
11. Mercer, R. E. and C. Di Marco: The importance of fine-grained cue phrases in scientific citations. In *Proceedings of the 16th Conference of the CSCSI/SCEIO* (AI'2003) (2003) 550–556
12. Myers, G.: *Writing biology.* University of Wisconsin Press (1991)
13. Reynar, J. C. and A. Ratnaparkhi: A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, D. C. (1997)
14. Teufel, S.: *Argumentative zoning: Information extraction from scientific articles.* Ph.D. thesis, University of Edinburgh (1999)

A Hedging Cues

Table 9. Base forms of all verb hedging cues used in the analysis.

appear	calculate	indicate	report	speculate
assume	estimate	note	see	suggest
attempt	imply	predict	seek	suspect
believe	indicate	propose	seem	

Table 10. All non-verb hedging cues used in the analysis.

about	essentially	partial	rarely
almost	evidently	partially	relatively
apparent	generally	possibility	slightly
apparently	likely	potentially	some
approximate	most	presumably	somewhat
approximately	mostly	probable	unlikely
around	normally	probably	usually
consistent	occasionally	quite	virtually

Each verbal cue given in Table 9 was expanded into four inflectional variants in the analysis presented in this paper: the base form, the third person singular present tense, the past tense, and the present participle form (e.g. appear, appears, appeared, and appearing, respectively).