2009

ANNOTATING A CORPUS OF BIOMEDICAL RESEARCH TEXTS: TWO MODELS OF RHETORICAL ANALYSIS

Barbara Ellen White

ANNOTATING A CORPUS OF BIOMEDICAL RESEARCH TEXTS: TWO MODELS OF RHETORICAL ANALYSIS

Spine title: Annotating a Corpus of Biomedical Research Texts

Thesis format: Monograph

by

Barbara Ellen White

Graduate Program in French Studies

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

ABSTRACT

Recent advances in the biomedical sciences have led to an enormous increase in the amount of research literature being published, most of it in electronic form; researchers are finding it difficult to keep up-to-date on all of the new developments in their fields. As a result there is a need to develop automated Text Mining tools to filter and organize data in a way which is useful to researchers. Human-annotated data are often used as the 'gold standard' to train such systems via machine learning methods.

This thesis reports on a project where three annotators applied two Models of rhetoric (argument) to a corpus of on-line biomedical research texts. How authors structure their argumentation and which rhetorical strategies they employ are key to how researchers present their experimental results; thus rhetorical analysis of a text could allow for the extraction of information which is pertinent for a particular researcher's purpose. The first Model stems from previous work in Computational Linguistics; it focuses on differentiating 'new' from 'old' information, and results from analysis of results. The second Model is based on Toulmin's argument structure (1958/2003); its main focus is to identify 'Claims' being made by the authors, but it also differentiates between internal and external evidence, as well as categories of explanation and implications of the current experiment.

In order to properly train automated systems, and as a gauge of the shared understanding of the argument scheme being applied, inter-annotator agreement should be relatively high. The results of this study show complete (three-way) inter-annotator agreement on

an average of 60.5% of the 400 sentences in the final corpus under Model 1, and 39.3% under Model 2. Analyses of the inter-annotator variation are done in order to examine in detail all of the factors involved; these include particular Model categories, individual annotator preferences, errors, and the corpus data itself. In order to reduce this inter-annotator variation, revisions to both Models are suggested; also it is recommended that in the future biomedical domain experts, possibly in tandem with experts in rhetoric, be used as annotators.

KEY WORDS: annotation, argument, biomedical text, computational linguistics, information extraction, rhetoric, text mining

ACKNOWLEDGMENTS

First I would like to thank my two supervisors, David Heap and Robert Mercer, who have been with me from the beginning of my graduate career to the completion of this thesis, sometimes in the background and other times in the front-line. I would not have made it through the seemingly endless process of putting this document together without their input and support. Secondly I thank my two annotators, JH and KP, who brought energy and commitment to this project, and without whom there would be no thesis. They gave me insight into the scientific, rather than the rhetorical or linguistic, point of view, and allowed me to see the process of annotating from a fresh perspective. A special thank-you goes to Heather Graves who took on the challenge of creating the second Model of argument applied in this study, and then consistently offered training and support to me and my annotators throughout our project.

I have received helpful administrative support from two of our Department Chairs, Jeff Tennant and Marilyn Randall, and I express my appreciation to both of them. I would also like to thank Brad Corbett of the Canada Research Data Centre at UWO for advice and feedback on my study design and statistical analyses. And I am especially grateful to Brian Hodgson who enabled me to defend this thesis sooner rather than later. Finally, my deepest thanks go to my late father, Frank Alexander White, for all his wisdom, integrity and caring; his spirit has never left me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

PREFACE

This thesis describes the methodology for and results of a project to annotate the argument structure (specifically the rhetorical strategies used by authors) of a corpus of biomedical research articles. The primary goal of this study is to compare and evaluate two different Models of argument by applying them to the same articles, using the same annotators. The secondary goal is to investigate the performance of annotators by having a lengthy training process, including feedback and discussions, as well as detailed analyses of the results by annotator. These two goals are achieved by quantifying the inter-annotator variation found in our results and identifying all sources of this variation. In addition to these two major goals I explore two related issues: the use of 'hedging' in biomedical writing, and the utility of developing a small set of 'Types' – canonical approaches to argument – only one of which would apply to each article in the corpus.

The motivation behind this study is the need to develop sophisticated automated search tools for biomedical researchers. Currently these researchers find it difficult or impossible to keep up-to-date with the enormous and rapidly growing volume of information being published on-line in their domains; there is a need to filter and constrain this flow, such that individual researchers are able to quickly access only the aspects of an article that are pertinent to them. In order to deal with this problem there is a need to develop tools that allow for automated Information Extraction (IE) from on-line research material; human-annotated data are the 'gold standard' for training such classifiers.

One approach to IE involves labelling parts of an article's text according to their roles in the authors' overall argumentative strategy. For example, a researcher might want to extract only the specific results of an experiment being described, filtering out background material and evidence external to the current study. The Models of argument applied in this project each contain a set of categories which represent the steps and strategies used by authors in developing and supporting an article's argument. Three annotators labelled units of text from our corpus by selecting one argument category for each unit from each of the two Models being tested. The results of this study show that some categories are clearly a greater source of inter-annotator variation than others; these data are crucial diagnostics for identifying weaknesses in both the Models applied here.

In order for data to be appropriate for machine learning algorithms, the level of inter-annotator variation on category identification must be kept relatively low. Thus we are looking for a Model of argument that balances simplicity – fewer categories mean fewer opportunities for annotators to disagree – with complexity – there are enough categories to differentiate between the types of information required by researchers. Ideally such a Model could be applied readily across different biomedical domains. Ultimately, for annotated data to be reliable and useful for researchers there is a need for a Model of argument that is relatively easy to understand and apply, matched with annotators who are comfortable with the corpus content and familiar with the concepts of argument and its structure.

In order to properly evaluate the Models of argument it is crucial to examine in detail all aspects of the inter-annotator agreement and variation found in the results from this project. Percentages of overall (here, three-way) inter-annotator agreement are necessary, but not sufficient, statistics for assessing the Models of argument; on their own they serve as guidelines rather than benchmarks. This is because there are a number of factors other than the Models themselves affecting the variation seen in our results. It is only through identifying the different sources of variation that recommendations can be made regarding improvements to the Models, and the annotation process itself. Thus I do not only provide overall statistics for inter-annotator agreement and variation, but I break down the variation by Model, category within each Model, annotator, and corpus article. For example, I not only give percentages for how often two particular annotators agreed across the corpus, but I also present crosstabulations showing how often these two annotators agreed/disagreed on each argument category, for both Models.

Since the 1990s the Computational Linguistics community has adopted the Kappa coefficient (Siegel & Castellan 1988) as the canonical measure of reliability in human annotation studies; it measures pair-wise agreement on categories between annotators while "correcting for expected chance agreement" (Carletta 1996: 252). More recently, however, questions have been raised about the appropriateness of this, and other, statistical measures, especially when applied to studies with more than two annotators, and results involving subjective judgements (Craggs & McGee Wood 2005; Artstein & Poesio 2008). The results of this project clearly show that much of the inter-annotator variation identified is not random ('by chance'): it reflects different annotator preferences

(one annotator is systematically inclined to select a category more often than the others) and the fact that under both Models the corpus data do not distribute equally into all categories (for example across articles and annotators, under Model 1 one category is chosen 36% of the time while another accounts for only 11% of the data). In addition, we may all legitimately disagree on a categorization (subjectivity), or there may be errors (annotations which clearly violate the specifications for a category). Thus, the Kappa coefficient cannot be applied to the results of this study; as I explain in detail in Chapter 5, part of future work will be the identification of the most appropriate statistical agreement coefficient.

Although the ultimate aim is to have data that are reliable enough to train automated systems, that is not the goal of the current project. Rather, given the goals stated above, I expect inter-annotator variation, and want to explore its dimensions. It is through examining and quantifying all sources of variation, as well as detailed analyses of the content of the individual corpus articles, that I am able to make strong recommendations regarding necessary revisions to the Models of argument, and the need for future annotators to be domain-experts. In addition, I stress that no further revisions should be made to the Models without input from biomedical researchers, the ultimate end-users of the IE tools that are to be developed.

Unique Contributions

The Argumentative Zoning (AZ) annotation scheme was identified as being applicable to "scientific text" (Teufel and Moens, 2000), but their corpus was composed of conference

papers in Computational Linguistics, a genre rarely similar to experimental biomedical research articles. Another previous approach, Zone Analysis (ZA), was based on AZ, but was applied to articles from four Biology journals (Mizuta et al., 2006). The current study, in contrast, covers articles from ten different biomedical domains, and hence gives a better picture of whether our two Models of argument are generalizable across disciplines. The ZA project involved only one annotator and thus there are no data on inter-annotator agreement - which is a crucial aspect of evaluating the utility of an annotation schema. In addition, in both of the above studies the entire article was annotated, a time-consuming process; I believe this creates unnecessary complexity. Given the focus of the research on rhetorical moves, in this project only the Discussion sections of articles were annotated: this is faster and leads to smaller, more manageable data sets while still covering the core argumentation of the paper.

Although the first Model of argument applied in this project is based on the above previous approaches of AZ and ZA, the second Model is based on Toulmin's argument structure (1958/2003), as adapted by Jenicek (2006) and Graves (2007). Although Toulmin's model has been applied in numerous other domains, this is the first time it has been used to analyze the argumentation found in biomedical research articles. Since AZ and ZA stem from relatively recent work in the fields of Computational Linguistics and Bioinformatics, whereas Toulmin's argument structure is based in classical approaches to logic and rhetoric, it is of interest to compare Models with such differing conceptual groundings to see which seemed better suited to describing the argument structures in our corpus data. Applying both these Models to all corpus articles allows me to evaluate the

strengths and weaknesses of each approach, and to provide significant insights into how authors argue by looking at the same articles from two different points of view. Ultimately the results of these comparisons will help in the improved design of future Models of argument, and will provide valuable data to the wider research community.

In any scheme where the sentence is used as the unit of annotation there are inevitably problems with complexity. Sentences which are grammatically complex frequently contain clauses which belong to different argument categories. But sentences with only one tensed verb may still be argumentatively complex i.e., they may seem to belong to more than one category at the same time. In order to deal with this problem, given that only one argument category is allowed for each unit, I develop a novel system of 'Trumping': Where annotators are conflicted in the face of complexity, the Trumping guidelines indicate which category Trumps another. In general, the category which prevails is the one which is most crucial to the authors' overall argumentative strategy.

One of the major contributions of the current project is the identification of the corpus articles themselves as a major source of variation. By giving overall (three-way) annotator agreement statistics under each Model of argument for each of the seventeen articles in the corpus, we see a striking range among them; previous studies have provided results for an entire corpus, thus implying a homogeneity across articles which may not have existed. Here it is revealed how overall averages can mask both very 'good' and very 'bad' levels of annotator agreement. In addition, by giving detailed discussions on fifteen of the seventeen articles, it is revealed how variation in writing style and skill,

as well as the level of technical terminology and complexity, affect the ease of annotation and the ability to agree, and are thus an important source of inter-annotator variation.

I discuss hedging from a theoretical point of view as an important aspect of the Rhetoric of Science, but I also exemplify how these strategies are implemented by presenting the distribution of lexical hedges in our study corpora by rhetorical category. These results show that the most frequently occurring hedges are consistent with those of a previous study with a similar corpus (Hyland 1998), thus suggesting commonality in hedging strategies across biomedical domains. Given the strong relationship found here between hedging and particular categories under both Models of argument, the most frequent lexical hedges could serve as cues for rhetorical functions/categories in future Information Extraction work.

The diagnostics and recommendations in this thesis will be used as crucial input to the "Interdisciplinary Approach to Text Annotation" project, currently supported by the Academic Development Fund (ADF) at the University of Western Ontario. This project will create an electronic corpus of biomedical research articles where the Discussion sections are annotated with tags representing argument categories, a resource which will be made available to researchers working in Computational Linguistics, Bioinformatics and Natural Language Processing. The information gleaned from the current study will inform both the Model of Argument being implemented and the choice of annotators in the ADF funded project. Work is currently being done in the Department of Computer Science at UWO by Mercer's group on extracting 'claims' from biomedical articles; the

results in this thesis from the application of the Toulmin-based Model will support this research.

In addition, the detailed results available on inter-annotator variation in this thesis will help to make researchers aware of the many different factors affecting agreement statistics, and the importance of carefully selecting the appropriate agreement coefficient when determining the reliability of human-annotated data. The extensive comparisons presented here between the applications of two conceptually different Models of argument will provide researchers with insight into how authors of biomedical articles argue, and into argumentation in general.

CHAPTER 1 INTRODUCTION

1.0 Background

The motivation for the study described in this thesis comes from the need for biomedical researchers to keep up with the extraordinary and on-going growth in the publication of academic articles: "Exponential growth of the peer-reviewed literature and the breakdown of disciplinary boundaries heralded by genome-scale instruments have made it harder than ever for scientists to find and assimilate all the publications relevant to their research." (Hunter and Cohen: 589) The identification of this problem led to the development of PubMed, a search tool developed by the National Center for Biotechnology Information at the U.S. National Library of Medicine which provides access to the MEDLINE database as well as other citation databases and journals. (Kumar and Vishnu: 13) MEDLINE contains bibliographic citations and abstracts in life and biomedical sciences; in 2005 an average of 1800 entries per day were added to MEDLINE. (Hunter and Cohen: 589)

Tools such as PubMed and Google Scholar can help to link researchers to relevant documents, but there is still a need for constraining information overload while keeping up-to-date on discoveries in one's field, as well as for finding creative ways to manage and learn from the most pertinent data. One approach has been the development of 'Text Mining' tools. Hearst defines Text Mining as "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources." (2003: 1) She states further that "The most active, and I think

promising, application area for text mining is in the biosciences." (2003: 3) Human-annotated corpora are often used as the 'gold standard' for training computers to perform automated Text Mining. Such annotation projects in biomedical fields have recently looked at domain-specific biological 'events' (Kim et al. 2008), classes of 'Gene Reference Into Functions' (Lu et al. 2006) and the classification of text fragments along dimensions such as 'focus' and 'polarity' (Wilbur et al. 2006).

One approach to Text Mining is 'Information Extraction' (IE) which "distills structured data or knowledge from unstructured text" (Mooney and Bunescu: 3). Tools are developed to extract only the particular portions or aspects of a corpus/an article – typically documents found on the internet – required for a specific purpose. In the biomedical domains most IE "efforts concentrate primarily on identifying bio-entities (mostly genes and proteins) and relationships among them" (Shatkay et al.: 2086). Another IE task is the development of systems to automatically resolve anaphora i.e., to identify the preceding text being referenced by a pronoun (e.g., *this, it*) or a phrase (e.g., *these factors*); this is a problem for training automated processors at a discourse level. Watters et al. used human annotators to perform anaphora resolution on a sample article in Biology, with the longer-term goal of building an annotated corpus. (2005)

Other approaches in IE at a discourse level involve annotating the rhetorical or argument structure of text; the current study falls in this category. This thesis reports on a pilot project where biomedical research texts were annotated under two Models of argumentation; the results will lead toward the development of improved Model(s), with

clearly defined categories, which will be readily understood and applied, and will hopefully minimize inter-annotator variation. The longer-term goal is to build an annotated corpus which will serve as training data for the development of automated IE tools for biomedical researchers. In this Chapter I present a brief introduction to Rhetoric in general, and the Rhetoric of Science in particular (Section 1.1), as well as descriptions of some previous annotation studies in argumentation (1.2). I then introduce the two Models of rhetoric that were applied in this study: Model 1 which grew out of earlier work which I describe in 1.2 (Section 1.3.1), and Model 2 which was developed by Dr. Heather Graves, based on Toulmin's theory of argument structure (1958/2003) (Section 1.3.2). In 1.4 I discuss 'hedging' in scientific writing and provide a list of the 'hedges' that were sought in and recorded from the study's corpora. Finally in Section 1.5 I introduce the notion of Argument Type, categorizations of typical macro-level arguments found in biomedical literature.

1.1 Rhetoric

The study of rhetoric[1], the art of using language so as to persuade or influence others, has a long history, beginning with the classical tradition of public speech. One of the seminal figures of the twentieth century in this field was Chaim Perelman; he had a background in the study of law and justice, and sought to investigate modern uses of rhetoric as it applied in a practical way in the society of post World War II Europe. Together with Olbrechts-Tyteca he published *La nouvelle rhétorique* in 1958. "Perelman came to regard

[1] In this thesis we will treat the terms 'rhetoric' and 'argumentation' as more or less equivalent. Although rhetoric is the art of persuasion whereas an argument is a logical structure - "an assertion supported by evidence" (Graves and Graves: 114) – which may use rhetorical techniques, in the field of Computational Linguistics, and in this study, this distinction is not critical.

rhetoric and dialectic as parts of a unified whole, in which dialectic functions as the theoretical underpinning for a theory of non-formal reasoning (argumentation), whereas rhetoric constitutes a practical discipline that utilizes dialectal techniques to convince or to persuade." (Gross & Dearin: 8) He saw Descartes' view of the universe, modelled on mathematical reasoning and requiring the elimination of the subjective, as inadequate for dealing with natural language and the real world. He stressed that all argumentative discourse is situated in a particular social and cultural context, crucially conditioned and affected by the audience ('universal' or 'particular') to whom it is addressed (Perelman 1971).

Following the work of the 'Informal Logic' movement, begun in the 1970's, van Eemeren and Grootendorst developed their 'Pragma-Dialectal' approach (1984). They believe the historically sharp distinction between rhetoric and dialectic is no longer relevant; rhetoric applies dialectic, and dialectic provides rhetoric with intellectual tools. Their descriptions of conventionalized 'argument schemes' and 'argumentation structures' (e.g., 'serial', 'convergent') offer a complementary alternative to formal logic models. Theorists differentiate the latter between 'structural' (logical) approaches - as manifested in the product of the reasoning process – and 'functional' (dialectical) approaches – which emphasize the process in which the structures arise, and the functions the argument structures fulfill.

1.1.1 The Rhetoric of Science

That scientific fields have their own particular rhetoric is a well-studied phenomenon, for example Bazerman (1988), Graves (2005), Gross (1990), Harris (1997), Locke (1992), Myers (1990). Scientific writing is sometimes regarded as being too straightforward to make use of rhetorical techniques in comparison to, for example, political oratory or legal arguing. As Bazerman puts it:

> ...scientific language is a particularly hard case for rhetoric, for sciences have the reputation for eschewing rhetoric and simply reporting natural fact that transcends symbolic trappings. Scientific writing is often treated apart from other forms of writing, as a special code privileged through its reliance on mathematics...(1988: 6)

And Perelman and Olbrechts-Tyteca state that "the use of conventionally admitted experimental and deductive techniques reduces, in science, the room for argumentation." (1969: 137) Closer examination of academic research articles and an understanding of scientific culture, however, reveal considerable use of rhetoric. Scientific research is an extremely competitive field, especially given the current situation described in Section 1.0 above, and scarce funding resources must be actively sought; scientists need to 'market' their results and expertise, and the use of rhetoric is one way to achieve this. Locke stresses that scientific experimentation and scientific writing are two distinct activities:

> To claim that the language of science is, or should be, objective because science itself is objective is to confuse cause and effect. The apparent objectivity is the result of the decision made about language usage, not the cause of it...Indeed, the repeatability of scientific results, the checking of results in other laboratories, is not so easily accomplished nor so unambiguous as the official rhetoric would suggest. But that is not the point. It is not that scientists misrepresent when, by use of the passive voice, they make the tacit claim that their work is infinitely repeatable, that every scientist who tries to duplicate it will observe precisely the same

> effects; it is that this methodological claim is made by way of rhetoric. (91)

Some common rhetorical techniques which I have observed in biomedical research articles: ensuring that you cite the appropriate colleagues; amassing sufficient evidence, from previous work as well as your own study, before making a claim; stressing the novelty and/or significance of your findings; hedging your statements where necessary. This last is an extremely important aspect of scientific argumentation, and one which will be discussed in detail in Section 1.4 below. It has perhaps become even more pertinent in the age of the internet: most (if not all) scientific publications are now available on-line, and many are Open Access, free to anyone; papers can thus be scrutinized by scientific peers anywhere in the world.

1.2 Previous annotation studies

Rhetorical Structure Theory (RST), a theory of text organization (Mann and Thompson 1988), was developed from a foundation in Halliday and Hasan's study of the 'cohesive' relations which connect parts of a discourse (1976): "It explains coherence by postulating a hierarchical, connected structure of texts, in which every part of a text has a role, a function to play, with respect to other parts of the text...[It also] provides a systematic way for an analyst (also called observer or judge) to annotate a text." (Taboada and Mann, 2006a: 425) (I note the use of the term 'rhetorical' rather than 'argument' structure; see footnote 1.) In the RST framework the network of 'local' rhetorical relations of a text are represented by means of a tree-like structure: each leaf is associated with a contiguous textual span (spans may not overlap); the internal nodes are labelled with the names of rhetorical relations e.g., 'Evidence', 'Elaboration'. These relations hold

between 'nuclei' and 'satellites': a satellite provides information thought to be believed by the reader, in order to buttress the information presented in the nucleus: the effect is that the reader's belief in the information presented in the nucleus is strengthened. (Marcu: 22)

In Computational Linguistics (CL) RST was originally used as a tool for text generation, but it has also been used for text summarization and discourse parsing (Marcu 2000). Probably the most ambitious metadata annotation project to date using RST was Carlson et al.'s corpus of 385 documents from the Penn Treebank (2001). Given the complexity of this task, the large number of rhetorical relations (78) and the hierarchical nature of the model, the results showed considerable inter-annotator variation. Although the authors acknowledge that some of this variation was the result of annotator errors, differences in "interpretation" were problematic:

> A larger issue though stems from variation in stylistic interpretation among annotators. The RST theory does not differentiate between different micro- and macro-levels of the discourse structure, and thus a fairly fine-grained set of relations operates at all levels. This, along with the concept of nuclearity, increased the variation in annotator interpretation. Even though we had very well-defined rules for segmenting the text into EDUs [Elementary Discourse Units], it proved quite difficult to make our already extensive guidelines more explicit in dictating how to assign nuclearity and relations. (108)

I note that their instruction document for annotators was very large, 87 pages in total: a 40 page manual plus 47 pages of appendices which summarize and index the manual.

In her 1999 PhD thesis Teufel introduced 'Argumentative Zoning' (AZ), a system for rhetorical analysis of scientific research articles. It was based on Swales' 'CARS'

(Creating a Research Space) model which describes a series of argumentative 'moves' that authors would use to convince their readers of the originality of their work. Her model was developed with the goal of creating automated IE tools for text summarization and improved citation indexes. Her interest in intellectual attribution was apparent in her taxonomy of argument categories: AIM, BACKGROUND, OWN, OTHER, BASIS, CONTRAST, TEXTUAL. Although she believed that the hierarchical structure of argumentation that RST uses is appropriate for representing fine-grained rhetorical relations, this level of detail is not relevant for tasks such as summarization, and AZ is not a hierarchical model. Unlike RST, which looks at the relationship between words/clauses and their adjacent texts, AZ is interested in the relationship between a sentence (their unit of annotation) and the 'macro' argumentative message (Teufel and Moens 1999, Teufel et al. 1999, Teufel and Moens 2000, Teufel and Moens 2002).

Three annotators, Teufel and two paid students with degrees in Cognitive Science and Speech Therapy, applied AZ to 25 conference articles in CL covering a wide range of different sub-domains. Note that Teufel chose not to declare herself the 'expert' annotator: "we believe that in subjective tasks like the one described here, there are no real experts." (Teufel and Moens 2002: 420) Written guidelines totalled seventeen pages, including a decision tree for category selection. Twenty hours of training were provided during which example papers were annotated and difficult cases discussed. Six weeks later six of the original articles were re-annotated. The results showed good stability and reproducibility, but this may have been due in part to the fact that overall 67% of sentences were annotated as OWN; without the need for finer-grained sub-divisions, this category would be easy to agree on.

Langer et al. (2004) developed an ontology of 'text types', more fine-grained than Teufel's, with sixteen different 'topics', not limited to aspects of rhetoric, such as 'data', 'problem'. Their schema is hierarchical with multiple levels; for example 'concepts' falls under 'framework', which in turn is a kind of 'evidence'. Two annotators who had been trained on an earlier version of their model worked independently on a corpus of on-line (German) research articles in linguistics. They were permitted to join or split sentences to create 'proper thematic units', and only one annotation per unit was permitted. Their aim was to go beyond automatic document classifiers to be able to categorize text segments; their preliminary results showed that some of the topic types were learned successfully under a classification algorithm, but others were problematic.

The 'Zone Analysis' (ZA) model was developed with the goal of extracting and organizing information related to experimental results in Biology articles (Mizuta and Collier 2004, Mizuta et al. 2004, Mizuta et al. 2006). They rejected the discourse-based concepts of RST and instead built their model on that of AZ, but developed a set of three groups of 'zone classes'. Group 1 includes Background, Problem-setting and OWN; the latter, in contrast to AZ, has five 'nested' sub-classes: Method, Result, Insight, Implication and Else (whatever does not fit in the previous four classes). Group 2 includes material which compares information in Group 1 to previous work: Connection and Difference; and Group 3 has only one class, Outline, which includes statements such as those referring to the section organization of the paper. Like AZ, ZA has a decision

tree for zone annotation, but given the model's hierarchical structure it is more complex than that for AZ; also, two of the branches allow text to be left unannotated.

There was only one annotator, the author Mizuta, and thus there are no inter-annotator agreement statistics. She annotated a corpus of twenty articles – they stress that only looking at abstracts misses crucial information – from journals in fields such as Molecular Biology and Cell Biology. They note the problem of complex sentences, where the difficulty in deciding on the most important or 'relevant' zone class can lead to increased inter-annotator variation; thus they allow units smaller than the sentence. They state: "a rhetorical status apparently corresponds to a proposition, the closest syntactic counterpart of which is a clause", thus the following sub-sentential units are allowed as constituents: coordinate and subordinate clauses, relative clauses, including participial versions, and infinitive clauses which provide the purpose expressed in the balance of the sentence e.g., [*To test...*] [*we performed...*] (Mizuta et al. 2006: 472-473)

The results of the annotations across entire articles show the most frequent classes are Result and Method with 30% and 28% respectively, followed by Background with 20%. Insight (authors' interpretations of their current data) and Implication (including limitations of their study and conjectures re future applications) (both under OWN) account for 11% and 9% respectively. (Mizuta et al. 2006: 482) On a subset of four articles in their corpus where they break down results by section, however, Implication accounts for 52% of the number of words in the Discussion section (Mizuta and Collier

2004: 1740). I mention this here as in this project we annotated only the Discussion sections of our corpus articles (see Section 2.1.1 below).

1.3 Models of argument used in this study

1.3.1 Model 1

Given the longer-term goal of training automated IE tools, and based on my previous research and annotation experiences with biomedical research articles, I wanted to develop a model of argument that was maximally simple, but still able to distinguish between the types of content that would be useful to researchers. RST is far too complex and fine-grained, and it did not seem to be a good match for biomedical corpus data. Langer's model also has too many categories, is hierarchical (with three levels below the top node of 'content') which adds to its complexity, and focuses more broadly on 'text types', not types of argumentation.

I had applied both AZ and ZA in my earlier annotations of biomedical research articles (White 2005a). I found that the category OWN in AZ (see 1.2 above) was too general to deal with biomedical research papers. This earlier corpus, as well as the corpora for the current project (see Sections 2.1 and 3.1 for details), contain articles reporting on biomedical experiments, and thus, unlike in the case of Teufel's CL conference papers, it seemed important to have more specific categories for the authors' experimental findings. Also both of the OWN and OTHER zones were for "neutral descriptions" (Teufel and Moens 2002: 416), and it was not clear to me exactly what that meant.

The ZA model breaks down the OWN category into classes that seemed to fit well with biomedical research papers, and their corpus is the only one of the four in 1.2 that annotated articles similar to those found in our corpora. On the other hand, the nesting of categories and the difficulties in accurately identifying sub-sentential units (see White 2005a for detailed analyses of my results) made ZA more complex than AZ.

In developing the first Model of argument to be applied in this project I have taken and adapted aspects of both AZ and ZA, incorporating the elements that seemed relevant to our corpora, and limiting the number of categories to five. Note that in both the AZ and ZA studies described in 1.2 above, and in my 2005a project, the entire articles were annotated; in the current study we would only be annotating the Discussion sections of our corpus articles (see Section 2.1). Thus it did not seem necessary to incorporate separate categories such as Problem-setting in ZA that would most commonly be found in the Introduction section. In addition I did not include the ZA Else class as I was aiming to develop a Model that did not require an 'other' category.

In the binary decision tree models for both AZ and ZA argument categories for a single annotation unit are eliminated one at a time – 'higher' classes win over 'lower'; neither allows for 'backtracking' and the ZA tree has two branches which allow text to be left unannotated. This approach did not match the thought processes involved in my previous annotations of biomedical texts, where I did need to backtrack, and where I did not approach each unit in isolation. In fact, I had found it extremely important to take into account the context of the surrounding text, both its content and its annotations, as well

as, on a more macro-level, of the overall argumentation and its structure. In addition, in this project my intention was that all text would be annotated. The preliminary Model 1 in Table 1 below is a non-ordered list – the numbers are simply nominal labels – of the categories available for each sentence

1) **PREVIOUS WORK/UNDISPUTED FACTS**
- Background to the current experiment
- Generally accepted knowledge in the field

2) **ISSUES UNDER DISPUTE**
- Context: different researchers have conflicting views regarding existing interpretations, significance, etc. of previous work
- Questions about which there are disagreements in the field
- Often includes motivation for current experiment

3) **METHOD/APPROACH/EXPERIMENTAL DESIGN**
- Basis for choice
- Descriptions of processes
- May include intermediate results (vs. (4))

4) **RESULTS OF CURRENT EXPERIMENT**
- Statements of findings
- Are they consistent with previous work?
- Do they contrast with previous work?

5) **ANALYSES/IMPLICATIONS OF CURRENT RESULTS**
- How they interpret their results
- Suggest why something did/did not happen
- Usually where the main claims are made
- Often are 'hedged' e.g., "These results may signify…"
- Plans/implications for future work, by authors or others

Table 1: Preliminary Model 1 categories

1.3.2 Model 2

1.3.2.1 Toulmin's theory of argument

Toulmin's work on informal logic (1958/2003) has been useful in a number of disciplines, especially rhetoric. His system of Claims, Warrants, Grounds, and Backing maps out a conceptual framework for argument that is easily accessible and adaptable to

a variety of situations. For example, textbook and other authors from areas such as technical writing (Graves and Graves 2007) and medical science writing (Jenicek 2006, see below) have adapted Toulmin's theory of informal argument to show students how to develop well supported and convincing arguments. In particular, researchers in science have found Toulmin useful for understanding and developing arguments. Toulmin's model has been adapted in creating software for argument analysis (Reed and Rowe 2006) and for use with Rhetorical Structure Theory (RST) to develop Explanation Generation Systems (Dalianis and Johannesson 1997). In the clinical model Toulmin's approach has been applied in designing a computational decision support schema (Fox and Modgil 2006), and in creating an "explanation framework" for an automated health care system (Shankar, Tu and Musen 2006).

1.3.2.2 Graves' adaptation of Toulmin

I had done some research into Toulmin's approach (*Uses of Argument*, 1958/2003), and was attracted by the apparent simplicity of his flowchart design which aims to combine logic with language as it is used in the real world. His examples, however, were from arguments in the legal field, such as questions of citizenship (1958: 105); it was not clear to me whether his model could be (readily) applied to biomedical research articles. Then I encountered the work of Jenicek (2006): he had adapted Toulmin to the health sciences, developing a flowchart (2006: SR30) as a guide to reading and writing the Discussion section of medical research articles. Jenicek's aim is to assist practitioners in "critical thinking" and the presentation of clear arguments, in order to be of practical use in the research community:

> Structuring the 'Discussion' section as a review of argumentation benefits more than the study and its authors. It allows the reader to grasp the real

> relevance and validity of the study and its usability for his or her decision-making in clinical and community care, research and health policies and program proposal, implementation, and evaluation. (2006: SR28)

I asked Dr. Heather Graves, a writer and rhetorician, and the Canadian expert on Toulmin, to look at Jenicek's model of argumentation to see if it might be appropriate for application to my corpus of biomedical research articles. Graves' insight was that since our task is to attempt to *analyze* existing arguments, rather than to create them, the model we need is essentially the inversion of Jenicek's. His flowchart begins with a research question or hypothesis ("Problem in context"), followed by internal and external evidence, then various types of limitation and qualification, leading to a "Claim". (2006: SR30) In adapting this to the annotation of biomedical articles, Graves' flowchart starts with the identification of a "Claim" and ends with "Problem in context" as the forward-looking implications of an experiment's results for the authors' field. Graves' 2007 adaptation of Jenicek in Table 2 below constitutes the preliminary Model 2:

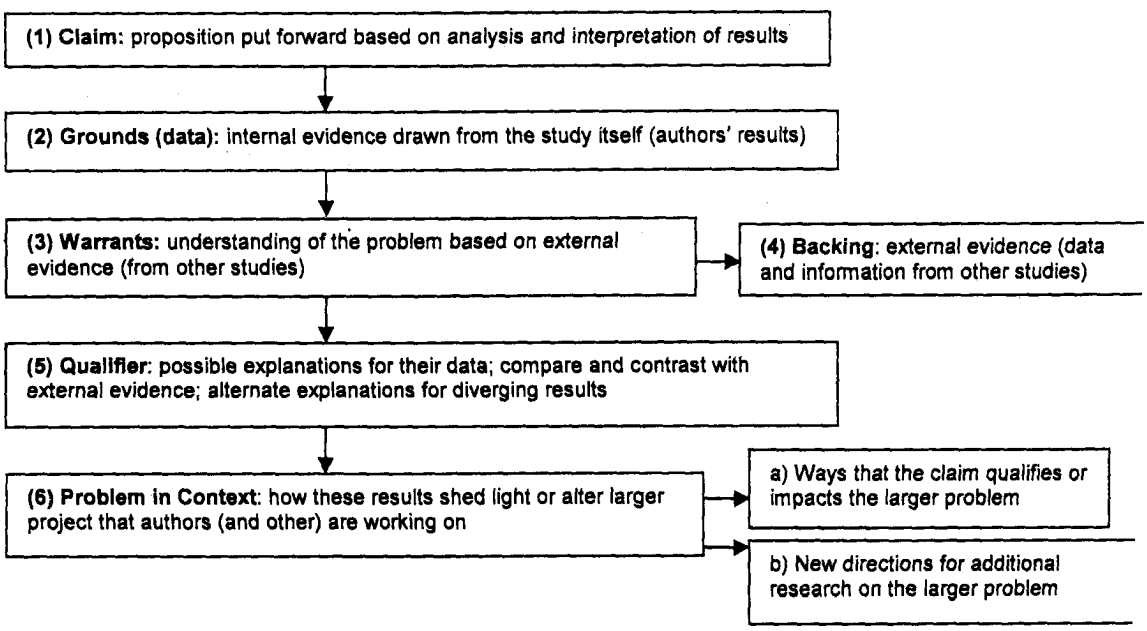

Table 2: Preliminary Model 2 categories

1.4 Hedging

Hedging is a pervasive and significant aspect of the Rhetoric of Science: since scientists' claims, when published, are open to criticism from their peers, it is crucial for authors of academic papers to 'hedge' or qualify their commitment to certain statements. Frequently there are multiple possible explanations for an experimental finding, and no scientist can have complete and up-to-date information in their field; scientific hubris could lead to professional embarrassment if you have made errors in your analyses or been overreaching in the scope of your experiment's implications for the field. As Hyland summarizes this:

> Hedges are important to scientists because...even the most assured scientific propositions have an inherently limited period of acceptance. Categorical assertions of truth in these circumstances are decidedly hazardous. Science writing involves weighing evidence, drawing conclusions from data, and stating circumstances which allow these conclusions to be accepted; it assesses the claims it makes. (1998:6)

In other words, hedging is inherent in the corpus data for this study, and is a critical component of how authors present their arguments. It thus seemed of interest to examine how hedges are used in our corpora in the context of the two different Models of argumentation being applied here. Under Model 1, one presumes hedging would be frequent in the Analysis category, but it would be interesting to observe the distribution of hedges in other categories. Under Model 2, it would seem that the notion of Qualifier, especially as it explores "possible explanations", is intrinsically linked to hedging. Hyland directly addresses Toulmin's argument layout:

> An examination of hedging can also contribute to our understanding of the practice of evidential reasoning and the structure of argumentation discussed by Toulmin (1958)...and others...Hedges clearly contribute to the repertoire of devices used to anticipate possible rebuttals, and their study can help reveal how writers move between grounds and claims in the process of gaining reader ratification for statements. (1996: 253)

Looked at from the opposite point of view i.e., in which categories are hedges *not* being used, the absence of hedges might indicate a particular category: Under Model 1, this might suggest category (1) – Previous work or undisputed facts, or category (4) – statements of Results of current experiment. Under Model 2, one supposes that Backing (4) and Grounds (2) would generally contain straightforward (unhedged) statements of observed results. In both of the models the rhetorical categories are reflecting (among other things) the difference between what is established as a fact in a particular biomedical field, and what is new and open to being challenged.

> In presenting claims, writers recognise that some ideas have been previously confirmed by the discourse community as constituting truth about the world. They can therefore assume that these statements form part of the set of background beliefs for the interpretation of evidence shared with the reader and that they can be relied on when presenting new hypotheses. Statements which employ such evidential support of hypotheses will therefore have a *factive* character. Every other statement by which the writer asserts the propositional content to be true as far as he or she knows and for which responsibility is assumed, is a hedged or *non-factive* statement. (Hyland 1998: 85)

The process of scientific inquiry over time thus consists of 'new' information becoming 'old', and some hedged Claims becoming Backing or Warrants for a future argument. For the purposes of this study, however, we were interested in capturing an argument as it exists at a moment in time on this transformational cycle, that is, at the date of publication of an article. Given the longer-term goal of developing IE tools, it was hoped

that the presence of certain hedges might serve as a cue to the identification of the argument category of a sentence.

It is worth noting here that previous data on the distribution of hedges seem to support the choice of Discussion section as the only part of our corpus articles to be annotated (see Section 2.1.1). It is generally acknowledged that it is this section of a scientific research article which contains its core argumentation: it is here that one moves from the Results section to what the results imply. In their article examining the function of the Discussion section in academic medical writing Skelton and Edwards note that it is in this section that authors present the significance of their data, and where it is appropriate for authors to speculate; they state "one can take the science out of rhetoric but not the rhetoric out of science." (2000: 1270) And this focus on speculation and interpretation implies that the majority of hedges should be found in the Discussion sections of academic research articles. In Hyland's corpus of 26 scientific research articles he found the "highest total frequencies of hedges and the highest density per thousand words" in the Discussion sections. (1998: 153)

1.4.1 Preliminary List of Hedges

In the interests of ease of investigation and with the longer-term goal of automated systems in mind, only lexical hedges were considered. Hyland presents three non-lexical or "discourse-based" hedging methods: "By referring to experimental weaknesses, inadequate knowledge, or limitations of the model, theory, or method used, writers can qualify commitment by offering a measure of propositional certainty." (1996; 272) In fact

these would be captured here by the choice of category at the sentence level: generally the Analysis category under Model 1 and the Qualifier category under Model 2. Use of the passive voice or impersonal constructions may be viewed as hedging by distancing the authors from the content of a statement, but these were not included in this study. Hyland notes that in scientific writing 85% of hedging is accomplished using lexical terms and only 15% by other strategies (1998: 104); also lexical items are easier to locate in text automatically. In addition, given our focus on rhetoric rather than the details of scientific content, we did not consider terms which hedge numeric quantities, such as *approximately* or *about.*

1.4.1.1 Modal Verbs

Modal auxiliary verbs in English are frequently used as hedges in academic writing; in contrast to declarative statements, they can express a degree of uncertainty, conditionality or subjectivity regarding what is being asserted. The six modal verbs chosen for this study were *could, may, might, should, would* and *can(not)* (see Table 3 below for a full listing of hedges). Based on my exposure to biomedical research articles, these are the most commonly used 'epistemic' modals. "Epistemic modality may be interpreted as truth with respect, not to worlds consistent with what is generally known, but to worlds consistent with what an individual speaker or hearer knows." (Cann 1993: 280) In other words, an author's opinion will be hedged in the context of the particular audience for whom they are writing, in our corpus a technically specialized one.

1.4.1.2 Non-modal Verbs

In addition to modals, verbs whose lexical content carries the notions of possibility or contingency are also used as hedges. These epistemic lexical verbs

> represent the most transparent means of coding the subjectivity of the epistemic source and are generally used to hedge either commitment or assertiveness. Their numerical significance [in Hyland's corpus of science articles] thus reflects their rhetorical versatility in contexts where categorical assertions rarely represent the most effective means of expression...Epistemic verbs therefore mark both the mode of knowing and its source (belief, deduction, report, perception) and thereby carry implications about the reliability of the knowledge itself. (Hyland 1998: 119-120)

These verbs can express probable inference (*suggest, indicate*), personal belief (*believe, think*), educated guesses regarding the future (*hope, speculate*), or an author's perception or opinion (*appear, seem*). In my experiences of researching and annotating biomedical articles I found the most frequently used epistemic lexical verbs were *suggest, indicate* and *seem*; the full list of these verbs is in Table 3 below.

1.4.1.3 Adjectives, Adverbs and Nouns

Although hedging is most often achieved in scientific research articles with the use of verbs, adjectives and adverbs are also employed to express probability (*probable, likely*), degrees of likelihood (*possible, perhaps*) or tentativeness (*preliminary*). In addition six nouns are included as possible hedges: five are nominal forms of some of the lexical hedging verbs (*appearance, assumption, belief, prediction, speculation*) and the sixth is the nominal form of the modal adjective *possible*. The full preliminary set of hedges is presented below:

MODAL AUXILIARIES	Could	ADJECTIVES/ ADVERBS	Probable/y
	May		Possible/y
	Might		Reasonable/y
	Should		Likely
	Would		Perhaps
	Can/Cannot		Maybe
LEXICAL VERBS	Appear		Preliminary
	Assume		Speculative
	Believe		Hopeful/ly
	Hope	NOUNS	Appearance
	Indicate		Assumption
	Predict		Belief
	Seem		Possibility
	Suggest		Prediction
	Think/thought		Speculation

Table 3: Preliminary lists of hedges

1.5. Scientific Argument Classification

1.5.1 Previous Approaches

In his study of the rhetoric of science Prelli finds that there are three "lines of argument", "structures of acceptable reasoning used over and over in scientific discussions." (1989: 216):

1. Problem-solution
2. Evaluative
3. Exemplary

These are based on Aristotle's *topoi* (topics) or categories of reasoning which Prelli maps onto current scientific discourse (185-207). To expand on the above he states:

> The general points I wish to make are (1) that there are in fact identifiable lines of thought that are used again and again in the sciences; (2) that these lines of thought legitimize scientific observations and claims because they derive from what is accepted and valued in scientific communities; and (3) that if we want to see what the logical formulas and characteristics of

> scientific discourse are, we must grant that these *topoi* identify structures of thought that scientists (and often others) find situationally reasonable. (1989: 216)

In her instruction guide for writing biomedical research papers Zeiger states that research papers fall into three different categories, depending on how they address the questions posed in the Introduction section. Writers will structure their argumentation according to which of the three they are presenting:

1. Hypothesis testing
2. Descriptive (e.g., a new structure and its implications)
3. Methods (advantages and disadvantages of a new method, its applications)

Although classifying entire articles, she acknowledges that the focus of the above are found in the Discussion section, where authors explain how their results support their position and fit within the knowledge base of their field. (2000: 176)

1.5.2 Argument Type

The concept of Argument Type was developed based on my experience over several years in reviewing and annotating biomedical research articles. I observed that the argumentation in these articles seemed to broadly fall into the three categories listed in Table 4 below:

1. **Novel**: Presentation of new procedure, application, results etc.
2. **Non-support/non-confirmation**: Their results are not consistent with work done previously in their field.
3. **Yes, but**: Their results show some confirmation of previous work, but: Some aspect is now in question (e.g., a new interpretation), or additional implications are presented ('not only').

Table 4: Preliminary Argument Types

Each biomedical article, when reviewed in its entirety, could thus be labelled with one of these Types (an assumption being that the core of its Argument Type would be found in the Discussion section). This categorization was seen as inherent to the data, and independent of whichever model of argumentation would be applied at the 'micro' (sentence or clause) level. It was hoped that the identification of an article's Argument Type could help to guide the choice of argument category at the micro level, especially in situations where a single unit seems to fit in more than one category. For example, given an article which has been assigned the Argument Type 2, if a sentence contains information on a previous study as well as current experimental results which are inconsistent with the earlier data, the argument category would relate to the current rather than previous findings.

1.5.3 Limitations

One of the problems with all three approaches above, however, is that the categories are extremely broad, and not necessarily mutually exclusive when it comes to biomedical research articles. For example, it is always the case that a research article is "novel" in some way, or it would not be published; thus the first in my list above in Table 4 may not be an informative or useful Type for argument categorization. Writers of biomedical research articles are typically presenting a blend of new and potentially challenging material (frequently hedged), while ensuring that their results are situated in the context of the existing knowledge base of their field. In other words, many if not most biomedical research papers will have some elements of all three of the above Argument Types. Most articles will have some aspects of both Prelli's first and second types of arguing, whereas

having the focus on analogy or metaphor found in his third type (205-207) is virtually non-existent in our corpus. And Zeiger's third category is already captured in the *BioMed Central* series of journals (see Section 2.1 for information on our corpus data) as they identify 'Methodology' and 'Research' articles as separate categories.

1.5.4 Further research on argument classification

I also surveyed a new (i.e., not previously seen or annotated) randomly selected set of ten articles from the BioMed Central database (see Section 2.1): my main goal was to attempt to find a set of argument 'templates' by examining in detail how each set of authors built their argumentation in the Discussion section. It was hoped that by looking at how the discourse is created i.e., how they present their evidence, how they attempt to deflect potential undercutters, etc. that a taxonomy of argument structures could be identified. This task was informed by previous models of argument in scientific papers (see Section 1.2), as well as the concept that in science "empirical inquiry proceeds under the assumption that knowledge is incomplete" (Sintonen: 122); thus I found data that fit categories such as 'Explanation', 'Implication', 'Incompleteness', 'Contradiction', etc. In addition to this 'bottom up' approach, I was also exploring how readily I could place each article in one of my three Argument Types above.

During this process I encountered again the difficulty in identifying argumentation when one is unfamiliar with the scientific content and the biomedical field (i.e., I am not the authors' intended audience); for example I was sometimes unsure whether statements of previous work were being presented as supportive or contradictory. When building from

a fine level up, outlines of argumentation became more and more complex; I was trying to find commonality, but instead I found a seemingly infinite amount of variation. It became clear that I could not find a set of argument templates: each article seemed to have a unique argumentative development and writing style. And at the macro level, I also found complexity: papers would not fit neatly into a single Argument Type, but often contained text belonging to more than one Type.

Given my above experiences as a linguist rather than a biomedical scientist, I hypothesized that those more knowledgeable about the biomedical domains being discussed in the corpus would be better equipped than I to assess the appropriateness and utility of the set of Argument Types I had developed. The notion of using a macro-argument categorization as a guide to annotation at a finer level still seemed a reasonable approach to explore. I would therefore ask annotators to test the application of Argument Types to the articles in our project's corpora, and seek their input regarding the refinement or complete revision of the list in Table 4.

CHAPTER 2 ANNOTATION OF TRAINING CORPUS

2.0 Introduction

This Chapter describes the training process for, and annotation of, a corpus of five biomedical research articles by three annotators, the project director (and thesis author) BW and two fourth-year undergraduate students in Medical Sciences. In phase I (Section 2.3.1) the two Models of rhetoric described in Chapter 1 were applied to two articles; units of annotation smaller than the sentence were allowed. In phase II (Section 2.3.2) three articles were annotated; here the unit of annotation was the sentence. For both phases I present results on inter-annotator agreement and disagreement, including a breakdown by argument category, under both Models, of sentences where we had three-way agreement. For phase II I also give a brief description of each of the three articles with example sentences that were found to be problematic during the annotation process. In addition I provide the distribution of hedges from the list in Chapter 1 (Table 3) across the five articles in the corpus. Following a summary of the difficulties encountered by annotators during the training process (Section 2.4) I describe revisions that were made to Models 1 and 2, the list of hedges, as well as the set of Argument Types (Section 2.5). Finally I summarize issues that remained after these revisions were made: problems inherent in the corpus data itself, and the question of the relationship between annotators with different skill sets and the biomedical corpus content (Section 2.6).

2.1 Corpora

All articles annotated during this project were randomly selected from BioMed Central, an on-line open-access publisher of research articles in science, technology and medicine.

It publishes a total of 198 peer-reviewed journals. (http://www.biomedcentral.com/info/about/whatis) Lists of the papers used in the training and final corpora are found in Appendices B and E. All articles are freely available from the www.biomedcentral.com site, and can be downloaded in text or PDF format (URLs are included in the Appendices). Although I wanted the corpora to represent a cross-section of biomedical fields, in the interests of some homogeneity, I have only selected articles from their *BMC*-series of journals. This series contains 61 journals in biological and clinical research; only articles describing experimental (i.e., non-clinical) studies were used in this project. (http://www.biomedcentral.com/info/authors/bmcseries?layout=printer#journallist) Because only Discussion sections were to be annotated, papers that combined Results/Discussion in one section were ruled out.

The Discussion sections of all five articles in the training corpus were colour highlighted according to the preliminary argument categories for both Models in Chapter 1. (Colour codes are not included in Tables 1 and 2, but they are included in the revised versions of Models 1 and 2 in Appendices C and D.) Hard copies of the colour-annotated articles were used in discussions during training as they allowed annotators to compare differences between annotators and Models in how they approached the text, especially when looking at the content and context of individual units.

The following conventions will be used for references to data from the corpora: All articles are referred to following the system of labelling found in Appendices B and E

i.e., T1-T5 for articles in the training corpus and C1-C12 for the final corpus. Specific sentences are referred to using a combination of paragraph number followed by sentence number e.g., 4-3 for the third sentence of paragraph four. All text that is part of a corpus article will be presented in italics.

2.1.1 Use of the Discussion section

As discussed above in Section 1.4 on hedging, the Discussion section of biomedical research articles is where authors go beyond what they have presented in their Results section to the potentially subjective interpretation and 'marketing' of their findings: this is where an article's core rhetoric is found. Some medical researchers such as Docherty and Smith see this use of techniques to 'win over' the reader as inappropriate, and the Discussion section as where this is most problematic; there:

> Authors may use extensive text without subheadings; expand reports with comment relating more to the generalities than to the specifics of the study; and introduce bias by emphasising the strengths of the study more than its weaknesses, reiterating selected results, and inflating the importance and generalisability of the findings. Commonly authors go beyond the evidence they have gathered and draw unjustified conclusions. (1999: 1224)

And Horton, as editor of *The Lancet*, sees the use of rhetoric, especially in the Discussion, as "manipulation by the author". (1995: 986)

But others, such as Skelton and Edwards in discussing the function of the Discussion section in academic medical writing, reject the above views, stating that rhetoric is both appropriate and useful in the Discussion section:

> A discussion cannot simply repeat the results as they seem beforehand or it is tautologous. In this sense every discussion is obliged to 'go beyond the evidence'. Every paper must reach a conclusion that is not contained in its results...In quantitative research, therefore, a central aim of discussions is to reinterpret the significant [in its statistical sense] as relevant – and that requires subjective interpretation of data. (2000: 1269)

In fact they stress that this subjectivity is a key part of the culture of science, as "a means of providing a context for the reader, of making science more than a list of facts or of numbers." (2000: 1269)

During my previous experience with *BMC*-series articles (White 2005a) I had annotated the entire article (excluding the abstract). This was extremely time-consuming, and I found that certain information was included more than once; for example, frequently findings from the Results section were repeated in the Discussion section, but with an interpretation of the result added. I have also noted that the Discussion section often includes background material such as previous studies or general knowledge in the field; thus, although the Introduction section also contains background, it is the information that is crucial to their argument that is found in the Discussion. In addition, biomedical researchers have told me that the Discussion is the key part of an article, and for many it is where they go directly after reading the abstract. Thus given the longer term goal of developing IE tools for researchers, I made the decision to annotate only the Discussion sections of the corpus articles throughout this project.

2.2 Annotators

In December 2007 I posted a "biomedical text annotator" position on the University of Western Ontario's "Work Study" job site. ("Work-Study" is a government-funded

program which pays full-time students for part-time work in the university environment.) The listing was aimed at senior Science or Medical Science undergraduate, or graduate, students. The requirements were: familiarity with biomedical research and academic writing; a good command of English; the ability to work independently and communicate results clearly; good attention to detail.

Two students applied, were interviewed, and were hired in January 2008. (A third student applied in mid-March, but the project was too far advanced to accommodate a new annotator.) Both were fourth (final) year undergraduate students in Medical Sciences: one female (KP) with a specialization in Physiology, and one male (JH) with specialization in Genetics and Biochemistry. They were both familiar with on-line resources for biomedical research such as BioMed Central and Pub-Med Central (the U.S. National Institutes of Health (NIH) free digital archive of biomedical and life sciences journal literature). As KP was hired several weeks before JH, they were initially 'out of phase' for the introductory training period, but by February they were working in parallel.

I (BW) was the third annotator throughout the project. During my previous annotation experiences it had become clear that my ability to analyze writers' argumentation was often seriously impaired by my lack of understanding of the scientific content of an article. I was not aware of generally accepted facts in a given field and thus could not always differentiate between statements of background and those discussing the current experiment. In addition, each biomedical subfield has a specialized technical lexicon with which I was unfamiliar. For example, I could not be sure whether a term referred to a

thing or a process. (These difficulties continued throughout my annotations in this project.) It was hoped that the senior science students would be better able to understand the technical content in the corpus, and thus to assess the rhetorical purpose and significance of authors' statements. Given our different academic backgrounds, one hypothesis was that JH and KP would be more likely to agree with each other on argument categorization, and to differ from BW's choice.

2.2.1 Orientation for annotators

Initially I provided JH and KP with a document which gave an overview of the annotation project and an introduction to what would be expected of them. I do not include the full document as it includes background material already provided here in Chapter 1. Below is the first paragraph:

> This project requires senior Medical Science or Science students to annotate the Discussion section of research articles from the BioMed Central *BMC*-series. This annotation will form a 'meta-data' layer on top of the existing text. (Note that no annotation will be required for charts, graphics etc.) I am interested in studying the 'rhetoric' or 'argumentation' used by writers of this type of paper: There is a substantial literature specifically devoted to the Rhetoric of Science; although 'hard' science has often been seen as presenting only facts (vs. e.g., the persuasive rhetoric of a politician trying to win our vote), those that study scientific papers, as well as senior biomedical researchers, agree that there is a great deal of 'marketing' of results and their implications in these data.

I then briefly described the motivation for and background to the project, and clarified why I was interested in annotators with their academic background:

> My experiences of rhetorical annotation have been hampered by the fact that I am a linguist rather than a biomedical scientist, and thus often have difficulty understanding the terminology and concepts being discussed in these data. One of the questions I want to explore is the difference in application of rhetorical Models between myself and 'experts' i.e., how much understanding of the

> detailed content is required to assess the authors' argumentation. I am also interested in comparing different annotations of the same article (each person will work in isolation) to see how much 'inter-annotator' variation occurs.

Finally I introduced them to how the annotations would proceed:

> Annotators will be provided with instructions and lists of rhetorical categories for each Model being applied. The annotation will be done either on hard copy with some form of coding e.g., coloured highlighting, or in electronic form (PDF or text) with highlighting tools. Although you may read an entire article, only the Discussion section will be annotated. (This is the section generally agreed to contain the core of the rhetoric.) One or more documents that are not part of the final corpus will be provided for practice; annotators will be free to ask questions and seek clarification during this process. When the annotations are being done, I will ask you to document your experiences e.g., how long did it take to perform? Where did you find difficulties and why? e.g., was it because you did not understand the Model being applied, or because you are not familiar with the scientific content? How do you think a given Model could be improved? In addition, other tasks may follow; these may include reviewing papers for content such as the use of 'hedging' cues e.g., "may", "suggest" or other linguistic features.

2.2.2 Training of annotators

Our annotation of the training corpus was divided into two phases: during phase I we annotated articles T1 and T2, and in phase II we annotated articles T3, T4 and T5 (see Appendix B). Throughout these two phases we gathered statistics on our annotations, discussed the Models and our processes. During this training period I had meetings with JH and KP (separately and together), and we were in regular email communication. In the early weeks I would review their annotations and see where they were not clear on certain aspects of the Models of argumentation we were applying; we would then discuss their category choices, and in some cases they made changes to their annotations. They were free to raise questions with me about either their understanding of the Models or their annotations of specific articles. Details of phases I and II are presented below in sections 2.3.1 and 2.3.2.

Although I had spent several years developing Model 1, at this initial training stage I was still relatively unfamiliar with Toulmin's concepts and Graves' Model 2, and had not yet attempted to apply it to any real data. I therefore had to undergo my own training under Graves to become comfortable with Model 2, both to be able to do my own annotating, and to supervise JH and KP in their applications of Model 2. This process is discussed below in section 2.2.2.1. Following this, my annotators and I met together with Graves so that JH and KP could clarify what types of data belong in each of her categories, give us feedback regarding some of their difficulties and offer suggestions for changes to her Model.

2.2.2.1 BW training on Model 2 under Graves

I had several meetings with Heather Graves to discuss how Toulmin's concepts of rhetoric might inform how I look at the articles in the corpus, and how I should understand the categories described in her Model. For example, she stressed the importance of unstated or implicit values and beliefs and suggested that I look at what they are, and how they are being appealed to, in my data. In order to understand the notion of Qualifier Graves explained to me that when a writer assesses the conditions under which an "unsympathetic reader" might refute a Claim, they will 'qualify' it to become more acceptable. Rhetorically, a writer counts on the "emotional investment of the reader"; when 'hedging' (e.g., the use of modal verbs such as *may*), the "onus is on the reader to confirm the results."

For my first test annotation, we agreed that there would be no constraints on the unit of annotation: I would simply change category wherever I felt there was a shift in rhetoric, thus allowing a phrase, clause or sentence to be a separate entity. Although the assumption is that statements are included because they are in some way relevant to the authors' argumentation, she advised me to experiment, and allow, if appropriate, text to remain un-annotated. Her overall theme was to be open to what I find in a given article, to "follow the rhetoric", rather than being driven by a computational approach (i.e., don't consider whether a machine could replicate what I am doing).

I first applied Graves' preliminary Model 2 (see Table 2, section 1.3.2) to article C4 of the final corpus (see Appendix E): "An informatics search for the low-molecular weight chromium-binding peptide" (www.biomedcentral.com/1472-6769/4/2). Following a discussion with Graves regarding this experience, and her review of my annotations, I went on to annotate two articles which were part of the training corpus (see Appendix B): T1, "Comparative 3-D Modeling of tmRNA" (www.biomedcentral.com/1471-2199/6/14) and T2, "Carvacrol and p-cymene inactivate Escherichia coli O157:H7 in apple juice" (www.biomedcentral.com/1471-2180/5/36) As with all other annotations, I read the abstract first, then the entire article, including any graphics or tables, and then annotated only the Discussion section (see 2.1.1 above).

As has been the case in my previous annotation experiences, I often found it difficult to evaluate the authors' rhetoric as I was not familiar with the scientific content of an article; this includes technical terminology (e.g., *D-stem*), acronyms (e.g., *Smpß, tmRNA*),

genre-specific lexical items (e.g., *trans-translation:* is the referent a result or a process?), and background material in the field. I was also learning about the Model 'on the fly', finding it especially difficult to understand what sometimes seemed to be subtle differences between categories e.g., Warrant vs. Backing. Throughout my annotating of the three articles above, I became aware that the rhetorically-based Model 2 seemed more conceptual than the informationally-based Models which I had applied in the past, and that an understanding of the authors' overall argumentative strategy was more critical under Model 2. For example, given a unit of text describing a finding from the researchers' study: Model 1 sees this as a piece of new information (Results of current experiment) whereas Model 2 sees it as internal evidence which is used to support a Claim (Grounds).

I found that I annotated virtually all of the texts, but that may have been because I was relatively unfamiliar with Model 2, and that my previous experiences required exhaustive annotation. (Later, with more experience under Model 2, and when a new 'external to the argument' category was added (see Section 2.5), I was better able to assess which units of text did not belong to one of the six preliminary categories). Interestingly, I found that I frequently left sentence-initial discourse connectives (e.g., *Therefore..., However...*) unannotated. (Graves describes these as part of the 'meta-discourse' rather than the argument.) Although they signal to the reader the relationship between the previous sentence (or a block of sentences) and what follows, I found they could also mark a change in rhetorical category e.g., a Warrant sentence followed by *However,* Claim; or a Grounds sentence followed by *Therefore,* Qualifier. In these instances, the pre and post

text received the appropriate annotation, but the connective, having functioned as a kind of 'switch' between them, was not really part of the scope of either. I also left some 'stop words' (e.g., *and, but*) unannotated, as well as some phrases such as *At this point* where they did not seem to have argumentative significance.

I note that the lack of constraint on the unit of annotation is liberating on the one hand, especially for complex sentences, but it can also be frustrating when it leads to an overly complex set of decisions: with more units there are more opportunities for inter-annotator variation. An example where sentence-splitting facilitated my annotation is in 3-5 from article T1: *The TLD mimicked the L-shape of canonical tRNA [39] | and | may be necessary for proper association* [sic] *tmRNA with the EF-Tu, SmpB, and subsequent binding to the ribosomal A-site* ("|" indicates a segmentation of the text marking a change in category). The first section is a Qualifier: it compares their current result with previous work in [39]; the second section I annotated as Problem in Context: their results shed light on future research. An example of a more problematic sentence subdivision occurred in 3-1 from article C4 is: *Although unexpected, | the results in this report | and | a critical review of other literature [9, 11-18], | suggest that an extracellular Model for Cr(III) biochemistry with respect to insulin signalling may be plausible (see Supporting Information).* The first phrase I had as a Qualifier, the second as Grounds (current results), the third as Warrant (overall view from external evidence), and the final section as a (crucial) Claim (a proposition put forward based on analysis and interpretation of results). Is the fact that their results were unexpected significant to their argument? Certainly it seems that combining the weight of their results with external evidence adds

to the force of their Claim. But given that the key purpose of this sentence is the Claim they are making, how important is it to have this fine-grained segmentation? This remains an open question in the field, and must ultimately be addressed in the context of a cost-benefit analysis (fine granularity and complexity vs. the simplicity of larger units and fewer opportunities for variation) for a given application.

In terms of understanding where to apply particular categories in Model 2, my greatest difficulties were in differentiating between: Warrant and Backing (the former is more conceptual); Grounds and Qualifier (is it a result, or a possible explanation for a result?); Qualifier and Problem in Context (are they explaining their results, or suggesting how things should be addressed in the future?).

2.3 Annotations of Training Corpus

2.3.1 Phase I: Annotation of articles T1 and T2

2.3.1.1 Model 1 and Argument Types

Since I was more familiar with Model 1, I presented my annotators with the preliminary version of Model 1 as in Table 1 in section 1.3.1. I told them that they would be annotating each unit of text with one of the five categories. I also introduced them to the notion of 'Argument Type', discussed in detail in Section 1.5. They were provided with the list of three original Types in section 1.5.2 (Table 4) and the below instructions regarding their application:

> This is the overall 'macro' argument categorization, and only one will apply to each paper. Base your choice on the core argumentation found in the Discussion section. Often the title of the paper will give you a good

> clue to this, but not always; e.g., a title may refer more to the process than the results. Always make this determination after reviewing, at a minimum, the entire Discussion section. These are mutually exclusive Types, so do your best to choose one of the three. If you feel none of the three applies, make a note of this, with what you feel might be an additional Type.

I then assigned the first two training articles: T1 from the *BMC Molecular Biology* series, and T2 from *BMC Microbiology* (see Appendix B). I decided to let their first annotation experience be relatively unconstrained; primarily I wanted them to become familiar with the Model, the concepts, and what it felt like to attempt to categorize real data. I did not specify a unit of annotation, allowing them to be free to segment the text as they felt it fell naturally into categories. Rather than instructing them on how to read the article (e.g., order of sections), I asked them to tell me what had worked best for them. It was made clear, however, that only the Discussion section should be annotated. I also allowed them to leave text unannotated, provided they noted on what grounds they had made this choice.

I also provided them with more detailed instructions regarding how I wanted them to approach the annotation task (this "Instructions to Annotators" document is found in Appendix A). I stressed that their training was to be an interactive period: I wanted them to make conscious their reading and annotation processes, and especially to let me know what they found easy or difficult. I asked them to keep a record of their experience: I wanted to know how they read the article: Was the abstract useful? Did they read every section, including charts and graphics? Did they follow the in-text citations to the bibliography? If they read the entire article, did they feel that the most significant rhetoric was found in the *Discussion* section? How familiar were they with the scientific content

of the article? If they had difficulty in categorizing a text, was this because of problems with the Model, or they were not clear on the meaning of the content? What was the most difficult part of this task for them? In addition, I asked them to keep track of the time they spent.

JH found difficulty in annotating sentences which seemed to combine experimental Results (category (4)) with the implications of the results (Analysis, category (5)). Interestingly he said that if he was looking for the results of an experiment he would first go to the Discussion rather than the Results section. JH noted that article T2 was "so well written" that he was able to readily grasp the ideas. However, in article T1, where he was focussing more on analyzing the individual sentences in the Discussion than the overall content, he told me: "If *understanding* the article was part of the tasks required of me, it was definitely the most difficult part." In T2 the hardest part for him was trying to find statements from category (2) (Issues under dispute); this was before he understood that he was not required to find any text that fit a particular category. (I note that this category was not in the final version of Model 1 (see Appendix C).) JH agreed that the most persuasive arguments are found in the Discussion section.

KP found sentences that required more than one category because they blended current conclusions and Analysis (category (5)) with material from Previous studies (category (1)). She found the most persuasive rhetoric in the Results section with its "more concrete" raw data, rather than the "stipulations" in the Discussion. In article T1 she had the greatest difficulty in "understanding it as [she] felt the author wrote it in a very run-on

jumbled way", and she was unfamiliar with some of the terminology. With T2 the hardest part was "understanding the background enough to annotate properly", although she found the description of their experiment to be clearly written. KP noted that the charts and tables in T2 were an important part of supporting the authors' argument.

2.3.1.2 Model 2

In early February I introduced my annotators to Model 2. I provided them with some brief background material on Toulmin and the development of Graves' 2007 Model. I explained that we believe that this type of rhetorical framework could be appropriate and useful in the analysis of biomedical research articles, and that our annotation project would collect data that could allow us to test this hypothesis by comparing Model 2 with Model 1. They were clearly more familiar with the concepts of Model 1 – it reflects to some degree the standard IMRD (Introduction-Methods-Results-Discussion) structure of the biomedical research articles they had been exposed to – but I instructed them to annotate the same articles (T1, T2), using the same methodology, under this new Model. In effect this practice annotation would serve as a tutorial on Model 2 by reviewing problems encountered in applying it to data with which they were already familiar. They were provided with the Graves' Model as presented in Table 2 in Section 1.3.2.2. The annotation instructions were parallel to those for Model 1: segment all text in the Discussion section into what they consider the most appropriate units of argumentation for Model 2; annotate each unit of text with one of the six possible categories, but if unsure, leave a text un-annotated, and include a note as to what difficulty they had encountered.

Both JH and KP found this Model much more difficult to understand and work with than Model 1. Some of their particular problem areas were in differentiating between Warrant (category (3)) and Backing (category (4)), Claim (category (1)) vs. a statement of current results (Grounds, category (2)), and understanding what a Qualifier (category (5)) does.

2.3.1.3 Results of annotations – articles T1 and T2

Given that units of annotation smaller than a sentence were allowed during this preliminary training, it was of interest to see how frequently this option was taken. With one exception (see below) annotators split sentences into only two units. Both T1 and T2 have 31 sentences in their Discussion sections. Of these 62 sentences, under Model 1 JH had 4 sentences with 2 categories each, BW had 11, and KP had 17 (she also had one sentence with 3 different categories). Under Model 2, JH had 7 'split' sentences, BW had 10, and KP had 6. On average 17.7% of all sentences were split under Model 1, and 12.3% under Model 2. These data are displayed in Table 5 below.[2]

	MODEL 1			MODEL 2		
	JH	**BW**	**KP**	**JH**	**BW**	**KP**
T1	2	8	11	4	6	1
T2	2	3	7*	3	4	5
TOTAL	4	11	18	7	10	6
% of all 62 sentences	6%	18%	29%	11%	16%	10%
Average	17.7%			12.3%		

Table 5: Number of 'split' sentences by Model and annotator

*Includes one 3-way split

[2] Given the preliminary nature of these annotations and the relatively low number of split sentences, data on whether annotators split a sentence into identical segments are not included.

Overall annotator agreement (i.e., where all three annotators chose the same category/categories for a given sentence) was higher under Model 1 (52% in T1 and 71% in T2) than under Model 2 (23% in T1 and 42% in T2); these results are presented in Table 6 below. It was not surprising to find better agreement under Model 1 since we were all more familiar and comfortable with its concepts than those of Model 2. Nor was it surprising that there was less agreement in article T1 than T2: we all agreed that T1 was difficult to read and understand (especially for BW for whom the content was quite inaccessible) whereas T2 was well-written, clear and more accessible for BW.

	Model 1		**Model 2**	
Article	**T1**	**T2**	**T1**	**T2**
# Sentences all agree	16	22	7	13
Total # sentences	31	31	31	31
Percent total agreement	**52%**	**71%**	**23%**	**42%**
Average for Model	**61.5%**		**32.5%**	

Table 6: Total annotator agreement by article and Model

In situations where all three annotators did not agree, there were four possible cases: we each chose a different category (all disagree), JH and KP agreed on a category, which differed from BW's choice (JK~B), JH and BW agreed vs. KP (JB~K) or BW and KP agreed vs. JH (BK~J). Under Model 1 all five instances of three-way disagreement occurred where one or two of us (but not all three) had split the sentence into smaller units; in three instances, however, we split the sentences, but still had overall agreement. The hypothesis that JH and KP would be more likely to agree and to differ from BW (see 2.2 above) was not proven true under Model 1: the largest two-way agreement was

between JH and BW (twelve of the nineteen cases), with only three JK~B and four BK~J sentences. It is worth noting that in eight of the twelve JB~K sentences, JH and BW had not split the sentence, but KP had. These data are summarized below in Table 7:

	T1		T2	
	No splits	**At least 1 split**	**No splits**	**At least 1 split**
All agree	15	1	20	2
All disagree	0	4	0	1
JK~B	0	3	0	0
JB~K	1	4	3	4
BK~J	1	2	1	0
Total	17	14	24	7
All sentences	31		31	

Table 7: Number of sentences in agreement/disagreement groups – Model 1

Under Model 2 we had a total of thirteen sentences with three-way disagreement, of which eleven occurred where at least one of us had split the sentence (eight where one annotator had split the sentence and the other two had not, and three where two of the three annotators had split the sentence). Unlike Model 1, here the two-way agreement categories were relatively equally distributed: eleven JK~B sentences, eight JB~K and ten BK~J sentences. Interestingly, in only five of these twenty-nine cases was there at least one annotator who split the sentence: three where the differing party split and those agreeing did not, and two where all three split the sentence. This is in contrast to Model 1 where thirteen of the nineteen two-way agreement sentences involved splitting. I do not know how to account for this difference; it is possible that as we were all less comfortable with Model 2, we were reluctant to split sentences and hence compound our uncertainty. The Model 2 data are presented below in Table 8:

	T1		T2	
	No splits	**At least 1 split**	**No splits**	**At least 1 split**
All agree	7	0	13	0
All disagree	1	8	1	3
JK~B	7	1	1	2
JB~K	1	1	6	0
BK~J	5	0	4	1
Total	21	10	25	6
All sentences	31		31	

Table 8: Number of sentences in agreement/disagreement groups – Model 2

2.3.1.3.1 Inter-annotator agreement on argument category

Under Model 1 most sentences with three-way agreement were in the categories Results (4) - fifteen sentences - or Analysis (5) - twelve sentences. The next most frequent category was Previous work (1) with seven sentences. This is partly a reflection of the fact that these three categories were generally the most commonly selected in all annotations across T1 and T2, not only those on which all annotators agreed. (Note that although detailed statistics of overall category distribution were not compiled during training, they are presented in results of the final corpus (see Tables 26 and 29 in Section 3.3.2).) We also all agreed on three sentences where we split into two distinct categories: Previous (1)| Analysis (5), Results (4)| Analysis (5) and Method (3)| Analysis (5). The complete agreement by category distribution for Model 1 is in Table 9 below:

	T1	T2	TOTAL
Previous (1)	3	4	7
Issues (2)	0	0	0
Method (3)	0	1	1
Results (4)	5	10	15
Analysis (5)	7	5	12
(1)|(5)	1	0	1
(4)|(5)	0	1	1
(3)|(5)	0	1	1
TOTAL	16	22	38

Table 9: Argument categories for sentences where all annotators agreed – Model 1

Since three-way agreement under Model 2 was slightly more than half that for Model 1 (see Table 6), there were fewer opportunities to have full agreement on individual categories. By far the most frequently agreed upon category was Grounds (2), accounting for thirteen of the total of twenty sentences (see Table 10 below). There were no instances of three-way agreement on a split sentence, or for the category Claim (1).

	T1	T2	TOTAL
Claim (1)	0	0	0
Grounds (2)	4	9	13
Warrant (3)	1	0	1
Backing (4)	1	1	2
Qualifier (5)	0	1	1
Problem (6)	1	2	3
TOTAL	7	13	20

Table 10: Argument categories for sentences where all annotators agreed – Model 2

2.3.1.3.2 Inter-annotator variation on argument category

Under Model 1, the most common variation was between categories (1) and (2) (Previous work vs. Issues under dispute), (1) and (4) (Previous work vs. current Results), and (4)

and (5) (current Results vs. Analysis of a result). Model 2 had one more category than Model 1 (six vs. five), and was also more difficult to apply, so there was more inter-annotator variation (although, being applied second, we were already somewhat familiar with the content of the articles). The main sources of disagreement were between Warrant (3) and Backing (4) (the two sub-categories of external evidence), Warrant (3) and Qualifier (5) (Is it external evidence, or a comparison between current and external evidence?), Qualifier (5) and Problem in context (6) (Are the authors' explanations discussing current findings, or are they focussed on the future?), Claim (1) and Grounds (2) (Is it a statement of result, or a Claim about a result?), and Claim (1) and Qualifier (5) (Is it an explanation for a result, or a Claim based on an analysis of results?). This final issue was one that continued through the project; the ability to differentiate between a Qualifier and a Claim seems to crucially require an understanding of both the scientific significance of a statement and the concepts of rhetoric that inform these two categories. (This issue is discussed more fully in Section 4.2.1.2.)

Given that Model 2 is crucially a Claims-based system, variation for this category (1) is of particular concern. As shown in Table 11 below there was considerable variation among annotators in the total number of Claims identified across T1 and T2: JH found sixteen, KP twelve, and BW only six. There were no instances where all three annotators agreed that a sentence, or a part of a sentence, was a Claim. Of the total of thirteen sentences where we had three-way disagreement, nine involved at least one annotator choosing Claim for all or part of a sentence. For example in the first sentence of article T1, JH chose (1), BW chose (2)-(1) (i.e., she split the sentence), and KP annotated as

Grounds (2); although there was clearly some 'overlap' here, this still counts as total disagreement since our choices were not identical. In T1, JH identified nine sentences and three sub-sentences as Claims (1), BW had two sentences and two sub-sentences, and KP had eight sentences as Claims. Far fewer Claims were identified in article T2: JH had two sentences and two sub-sentential units; BW had one sentence and one sub-sentence; KP had 4 sub-sentences as Claims. It is possible that BW was not recognizing Claims as she was unfamiliar with the biomedical fields, but it is also possible that KP and JH were over-identifying Claims as they were unfamiliar with what constitutes a Claim (see above).

	JH		BW		KP	
	Single	**In split**	**Single**	**In split**	**Single**	**In split**
T1	9	3	2	2	8	0
T2	2	2	1	1	0	4
Total	11	5	3	3	8	4
All Claims	16		6		12	

Table 11: Number of Claims per annotator, in single or split sentences

2.3.1.4 Summary of Phase I

Overall JH and KP were more comfortable with the scientific content (terminologies. methodologies, etc.) of these articles than BW, although they were not necessarily familiar with specific details. BW required more time re-reading an article e.g., in T2, trying to ascertain whether *yeasts* and *spoilage yeasts* had the same referent.

Although overall inter-annotator agreement was relatively low (especially under Model 2, see Table 6 above), this preliminary exposure to both the Models, as well as the process of annotating, was extremely productive. Questions such as whether to adapt or eliminate category (2) (Issues under dispute) in Model 1 were raised, and major issues such as clarifying the definition of a Claim (1) (Model 2) were introduced. I was able to exemplify for them how the differing concepts behind the two Models apply in our data. For example, in sentence 3-3 of T2, the authors state *However, this is the first report of the successful application of...* (my underlining for emphasis): under Model 1 this sentence is a statement (one of many) of their current Results (4), but under Model 2, its particular rhetorical significance (a proposition that no one else has been successful with this before) can be brought to the fore by labelling it as a Claim (1) rather than Grounds (2).

I had expected more sentence 'splitting' than was found in the results of our annotations (see Table 5 above). Where sentence fragmentation did exist, we were unlikely to have three-way agreement on the entire sentence: under Model 1 it accounted for only three of the 38 sentences on which we had total agreement, and under Model 2 there were no split sentences on which we had overall agreement (see Tables 7 and 8 above). Ultimately I decided that the problems with allowing units of annotation smaller than the sentence (we did not necessarily agree on what these units were, it increased the possibilities for inter-annotator variation) were too great; henceforth we would use the sentence as unit of text in all annotations. Although I believe that the clause is the ideal unit of annotation to capture the subtleties of argumentation, it was clear that KP and JH were not totally

comfortable with identifying clauses, and there was not time in this project to deal with this level of complexity. In terms of longer range IE tools, the sentence remains the easiest unequivocal unit of text to identify automatically. The question of what is the best unit of annotation for argument analysis is addressed further in Section 5.2.

2.3.2 Phase II: Annotation of articles T3, T4 and T5

Next we (BW, KP, JH) met with Heather Graves to discuss her Model. The main goals were to ask questions related to problems we had encountered during our annotations of articles T1 and T2, and to the theoretical concepts behind Toulmin's Model of argumentation; ultimately I hoped to establish a shared understanding of Graves' adaptation of Toulmin, and to clarify the definitions of each of her categories. The key findings from this consultation were that: a Claim (1) can be either 'original' or 'derived' (i.e., based on current findings only, or on external evidence); this type of academic biomedical writing tends to have more 'hedging' of Claims than other genres; a Qualifier (5) will often follow a Claim, in order to deflect anticipated criticism; a Warrant (3) can be an assumption or explanation which supports a Claim ("understanding of the problem").

Following the above discussion with Graves, I instructed KP and JH to annotate the next three articles from the Training Corpus (see Appendix B): T3 from *BMC Chemical Biology*, T4 from *BMC Medical Genetics*, and T5 from *BMC Cell Biology*. As with T1 and T2 the task was to colour code the Discussion section of each article in electronic format (using WORD), but unlike T1 and T2, they were to use the sentence as unit of

annotation: no 'splitting' of sentences was allowed. I also asked them to keep notes regarding areas where they felt changes to either of the Models would be helpful for future annotating. When all three annotators had completed the three articles, JH was to compile statistics on our inter-annotator variation under Model 1, and KP under Model 2. Below are selections from the instructions which I sent to them on February 25, 2008:

> Below are some general points to bear in mind as you annotate these articles:
>
> 1) When selecting a category, remember that I am always most interested in the RHETORIC/ARGUMENTATION i.e., we are trying to assess what the authors' underlying rhetorical purpose is (more so than just the informational content). Thus, if a sentence contains a result in the first part, followed by a "suggests that...", in general the main rhetorical purpose of the statement is likely to be what follows "that...": e.g., Implication of the result, a Claim, etc.
>
> 2) We are always evaluating only what is in the text of a given paper, not the science, choice of experimental methodology, etc. These things may make for a more weakly presented argument, but we are only identifying the different components of their argument, not critiquing it.
>
> 3) For categories such as Issues under dispute (2) in Model 1, bear in mind that this refers only to disputes that are presented 'on the page', i.e., there may be disagreements in the field at large, but unless they are discussed specifically in the article, they cannot be classed as such.

2.3.2.1 Introduction to Hedging for annotators

Although the major focus of training for the annotators was familiarity with the two Models of argumentation and the annotation process, I also introduced them to the notion of using hedges in scientific writing. I have provided more in-depth background on the concepts of hedging in scientific literature in Section 1.4; here I present a brief record of how the student annotators were introduced to hedging. After they had preliminary experience with annotating articles T1 and T2 from the training corpus under both

Models, I provided them with some brief background material related to hedging, including the below definition from Hyland:

> A hedge is therefore any linguistic means used to indicate either a) a lack of complete commitment to the truth of a proposition, or b) a desire not to express that commitment categorically. Hedges express tentativeness and possibility in communication, and their appropriate use in scientific discourse is vital. Hedging enables writers to express a perspective on their statements, to present unproven claims with caution, and to enter into a dialogue with their audiences. (1996: 251)

I informed them that hedging is a key strategy in presenting results of scientific experiments: writers must always be aware of the limitations of their work, the possible responses of other scientists, the risks of hubris in their community, and the recognition of the inherent 'incompleteness' in their part in the wider scientific investigations taking place in their field. Based on their own readings as undergraduate students, as well as their annotation experiences with the *BioMed* training articles, they were familiar with the most common hedging strategies e.g., writers using verb forms such as *suggests that* or *seems that*, adjectives such as *possible* or *probable*, or modal verbs such as *may* and *might*, to 'soften' the interpretations of their findings and their recommendations. They understood that this argument 'etiquette' is important in the scientific research community.

The annotators were told that part of the annotation project was to explore the relationship between hedging and rhetoric, and thus I wanted to examine the frequency and distribution of hedges in our corpus. In particular, I wanted to see if there is a pattern of hedging in the Discussion sections that relates to particular categories of argumentation. Since our two Models have different theoretical foundations, it would be

of interest to see where hedges appear in the different argument structures. I asked them to record instances of hedges in each of the training articles T1-T5 from the list in Table 3. Because at this stage they were just becoming familiar with the Models and our inter-annotator variation was high, we would not be able to extract a clear relationship between argument category and hedge: for example, sentence 4-1 in T4 contained the hedge *could*; under Model 1 it was annotated as Previous work (1) by JH, Results (4) by BW and Analysis (5) by KP. It was hoped that in the final corpus there would be improved inter-annotator agreement such that meaningful relations regarding hedges could be inferred e.g., *suggest* might generally indicate a category of Claim (1) under Model 2. Data on hedge distribution were compiled for all five training articles together; these results will be presented in Section 2.4.2 below.

2.3.2.2 Results of annotations of T3, T4, T5

The three-way agreement statistics for articles T3, T4 and T5 are presented in Table 12 below. The most striking result of the phase II annotations is that the average total (three-way) annotator agreement is virtually identical for each Model to that of phase I: Model 1 had 61.0% here and 61.5% in phase I, and Model 2 had 32.0% here and 32.5% in phase I (see Table 6 above). This is particularly noteworthy given that the Discussion sections of articles T3-T5, unlike T1-T2 (31 sentences each), are of very different lengths (ranging from 16-42 sentences), each of the five articles is from a different biomedical field (see Appendix B), and in phase II, sentence-splitting was not allowed.

	Model 1			Model 2		
Article	**T3**	**T4**	**T5**	**T3**	**T4**	**T5**
# Sentences all agree	21	10	19	15	6	6
Total # sentences	42	16	27	42	16	27
Percent total agreement	50%	63%	70%	36%	38%	22%
Average for Model	61.0%			32.0%		

Table 12: Total annotator agreement by article and Model

Data on the distribution of agreement/disagreement groupings for articles T3-T5 are presented below for Model 1 (Table 13) and Model 2 (Table 14). These patterns are extremely similar to those in phase I under both Models 1 and 2 (see Tables 7 and 8) with one notable exception: Under both Models the two-way JK~B agreement forms a larger percentage of the total two-way agreement than in phase I: 39% (13/33) here vs. 16% (3/19) in phase I under Model 1, and 48% (23/48) vs. 38% (11/29) under Model 2. This variation is largely a result of the particular skewing toward JK~B in article T3 which will be discussed in detail in Section 2.3.2.2.1 below:

	T3	T4	T5	TOTAL
All agree	21	10	19	50
All disagree	1	1	0	2
JK~B	11	1	1	13
JB~K	7	3	6	16
BK~J	2	1	1	4
TOTAL	42	16	27	85

Table 13: Number of sentences in agreement/disagreement groups – Model 1

	T3	T4	T5	TOTAL
All agree	15	6	6	27
All disagree	6	1	3	10
JK~B	13	3	7	23
JB~K	4	3	4	11
BK~J	4	3	7	14
TOTAL	42	16	27	85

Table 14: Number of sentences in agreement/disagreement groups – Model 2

Breakdowns by argument category for sentences with three-way agreement are given below for Model 1 (Table 15) and Model 2 (Table16). As in phase I (see Table 9), under Model 1 categories Results (4) and Analysis (5) are strongly represented accounting for 58% (29/50) of the total. Unlike phase I, however, category (1) (Previous work) accounts for 34% (17/50) of the total vs. only 18% (7/38) in articles T1 and T2. It is hard to say how much of this difference is due to increased familiarity with the category/Model, or the nature of the content of the particular articles.

	T3	T4	T5	TOTAL
Previous (1)	5	5	7	17
Issues (2)	1	0	0	1
Method (3)	0	2	1	3
Results (4)	10	2	6	18
Analysis (5)	5	1	5	11
TOTAL	21	10	19	50

Table 15: Number of sentences where all annotators agreed by category – Model 1

Under Model 2 we only agreed on a single Claim (1) across the three articles, but in phase I we had not agreed on any Claim (see Table 10). Category (2) (Grounds) still accounted for the majority of three-way agreement sentences, but at a reduced percentage: 37% (10/27 sentences) vs. 65% (13/20) in phase I. This was balanced by an

increase in the representation of sentences in the categories Warrant (3), Backing (4) and Qualifier (5). As above, the small sample size plus the fact that we were in a training phase make it difficult to definitively account for this variation.

	T3	T4	T5	TOTAL
Claim (1)	0	1	0	1
Grounds (2)	7	1	2	10
Warrant (3)	3	1	1	5
Backing (4)	0	2	3	5
Qualifier (5)	4	1	0	5
Problem (6)	1	0	0	1
TOTAL	15	6	6	27

Table 16: Number of sentences where all annotators agreed by category – Model 2

Despite the increased corpus size in phase II – 85 sentences vs. 62 in phase I – the total number and distribution of Claims remained remarkably similar. In fact there were more Claims in phase I (34, see Table 11) than the 30 found in phase II (see Table 17 below). It is possible that the change away from allowing units smaller than the sentence may have had some effect on this, as in phase I this increased the total number of possible annotation units beyond 62. Here JH had thirteen Claims, BW six and KP eleven; in phase I we had sixteen, six and twelve respectively. As discussed in Section 2.3.1.3.2 for phase I, it is not clear what accounts for this JH and KP vs. BW variation.

	JH	BW	KP	TOTAL
T3	5	3	5	13
T4	1	2	1	4
T5	7	1	5	13
TOTAL	13	6	11	30

Table 17: Number of Claims per annotator

2.3.2.2.1 Article T3, from *BMC Chemical Biology*

This was the most difficult of the three articles in this training set: the content was technically complex, the writing was often confusing, and it was the longest in terms of number of sentences (42 in the Discussion section). The overall inter-annotator agreement was generally poor, but somewhat better under Model 1 (50%) than under Model 2 (36%) (see Table 12).

Under Model 1, overall agreement was good in the first two paragraphs (twelve out of sixteen sentences), but the balance of the article was problematic. Paragraphs three and four (totalling seventeen sentences) contain a dense presentation of previous studies, speculations regarding some of these, statements of their current results, and discussions as to whether these are consistent or in contrast with previous work. The final paragraph discusses how and why their current model is a return to an earlier one which had been abandoned. BW annotated all twelve sentences of paragraph three as Results (4) – either current findings, or statements putting these in the context of previous work. However, JH and KP frequently chose either Previous work (1) or Issues under dispute (2). These findings brought to light two problems with Model 1: difficulty distinguishing between (2) and (4) for statements of conflicting data, and difficulty choosing between (1) and (4) for statements relating to studies done before the current experiment. In the final paragraph the main variation was between Issues under dispute (2) and Analysis (5): is it disagreement in the field, or an analysis of why this disagreement exists? Again, the category Issues under Dispute (2) seemed to be problematic.

There was only one sentence (2-7) under Model 1 where we all disagreed: *When binding is achieved at very high ribozyme concentrations, the ribozyme-substrate complex is sterically hindered to such an extent that the complex is locked into a poorly active conformation.* As there was no citation, BW believed it related to their current Results (4), but KP saw it as a generally accepted fact i.e., Previous (1); JH, however, focussed on the explanatory aspect, and annotated it as Analysis (5). This is one of many examples where BW's lack of background in biomedical sciences hindered her in being able to choose the appropriate argumentative category: not knowing the field – either the terminology or the background material – she could not recognize what was being discussed. Thus the extremely technical content of this paper created difficulties for BW and led to the largest 'two-way' agreement being between JH and KP, both of whom have an academic background in medical science: There were eleven sentences where JH and KP were in agreement against BW (JK~B), seven where JH and BW agreed against KP (JB~K), and only two where BW and KP agreed against JH (BK~J) (see Table 13).

Although we had total agreement on only 15 of the 42 sentences under Model 2, there was total disagreement on only six sentences (see Table 14). In four of these six, KP (only) had selected category (6) (Problem in context); in at least two of these (3-6 and 5-5), the choice of this category was an error. This category, as represented in the flowchart in Table 2, focuses on how the authors' current results "shed light on" a larger research issue; in other words, this is a concluding, and forward-looking, category, not one for presenting existing external evidence. Both of these sentences use the past tense, and are clearly referring to previous work. In sentence 1-5, BW realized in retrospect that her

choice of category (6) was also an error. After listing previous results, the authors state: *The results of the present study are in accord with these conclusions*; this should have been annotated instead as Qualifier (5), "compare and contrast with external evidence". JH and KP's choice of Grounds (2) for this sentence did not capture its rhetorical significance: it was not presenting their results, but crucially stating that their results are backed up by numerous other researchers. In fact JH and KP agreed on thirteen sentences where BW had a different category, an even larger number than under Model 1. And as with Model 1, most of these JK~B sentences occurred in paragraph three, and stemmed from BW mistakenly thinking statements refer to external evidence, when they actually relate to current results. There were only four each of JB~K and BK~J sentences.

JH and KP each identified five Claims (three of which they agreed on), and BW found three, none of which overlapped with theirs (see Table 16). Of the three JK~B Claims, BW felt in retrospect she might change from Grounds (2) to Claim (1) on two. BW's three Claims were all in the final paragraph, one which was even more difficult to annotate under Model 2 than Model 1 (see above), and where most of the total disagreement occurred. One of these three-way disagreement sentences (3-3) was a classically problematic sentence: *Since the...are composed of DNA, the four remaining ...were the critical residues for stabilisation.* For BW this was a statement of Grounds (2); for KP it was an explanation, so Qualifier (5); and for JH it was a Claim (1). All three categories could potentially apply to sentence 3-3. One of KP's choices of Claim, however, was an error: sentence 1-3 begins with a cited previous work, then states *...and <u>we postulated</u> that a similar change <u>was</u> required for...[20].* (my underlining) Although

this is work done by two of the present authors, it was done eight years earlier; the past tense of the verbs confirms that it was a *previous* Claim, and thus should be annotated, as confirmed by Graves, as a Warrant (3), not a Claim. (Note, however, that a *current* Claim (1) may be based, at least in part, on external evidence.)

2.3.2.2.2 Article T4, from *BMC Medical Genetics*

Of the three articles in phase II this was the shortest (16 sentences in the Discussion section), and the content was the most accessible for BW: although its topic was gene polymorphisms it related these to various human populations. Total overall annotator agreement was higher under Model 1 (63%) than Model 2 (38%) (Table 12).

There was only one sentence (4-1) under Model 1 where we all disagreed: *Similar observations could be made for the reported association of...in a* [sic] *White and African American populations from United States* [sic], *which we failed to replicate [30].* This first sentence of the fourth paragraph carries over from the previous sentences (*Similar observations*); the contrast with previous work led BW to choose Results (4), but KP saw it as interpretation of their results i.e., Analysis (5). JH's choice of Previous (1) was not correct: although this sentence refers to previous work, it is a crucial aspect of their current argumentation, not background to their experiment. The inter-annotator agreement was not biased toward the two science students: there was only one sentence where JH and KP agreed vs. BW's choice (JK~B), one sentence where BW and KP agreed vs. JH (BK~J), but three sentences where JH and BW agreed vs. KP (JB~K) (see Table 13). Variation found between Previous work (1) and Analysis (5) brought up a

difficulty with this Model where context is a key factor, for example in sentence 5-4: *The failure to replicate...is a common event in the search for genetic determinants of complex diseases, due either to genuine population heterogeneity or a different sort of bias [33].* Here the authors are making a general statement (*The* vs. 'Our'); KP chose (1) as it cited a previous result, but JH and BW chose Analysis (5) as it discussed possible general reasons for failing to replicate, and thus (indirectly) related to their current results. Looking at this sentence in isolation, one might choose Previous(1), but in the context of the previous sentences, it is clearly crucial support for the authors' findings; Model 2's notion of using external evidence to support a Claim seems more appropriate for this type of sentence.

It was the sentence just discussed above (5-4) that was the sole sentence where all annotators disagreed under Model 2. JH had Qualifier (5) – a "possible explanation for their data", BW had Warrant (3) – "understanding of the problem based on external evidence", and KP had Problem in context (6). This latter was based on a misunderstanding of category (6): as in T3 above, this should only include sentences that are 'forward-looking' i.e., how the current results "shed light on" the future of the field. Both (5) and (3), however, seem to be legitimate choices. For the sentence (4-1) with total disagreement under Model 1 (see above), BW and KP agreed on Warrant (3) under Model 2; JH had Grounds (2), but following a review of our annotations, he said that he would change his to Warrant (3). JH and KP each identified one Claim and BW had two; we all agreed that sentence 3-2 was a Claim. The two-way agreement was split evenly: three sentences each for JK~B, BK~J and JB~K (see Table 14).

2.3.2.2.3 Article T5, from *BMC Cell Biology*

The total overall agreement was strikingly better here under Model 1 (70%) than under Model 2 (22%) (see Table 12). There are 27 sentences in the Discussion section. BW found the material fairly difficult to understand and the terminology confusing, even though she had read the entire article before annotating. It seems the categorization under Model 1 was more straightforward than Model 2: many sentences fit clearly as Previous (1), Results (4) or Analysis (5). Under Model 2, however, variation between Warrant (3) and Backing (4), and Claim (1) and other categories, were major factors in the low overall agreement (see below).

There was no sentence under Model 1 where we all disagreed. Of the eight sentences with two-way agreement, there were one each with JK~B and BK~J, and six with JB~K (see Table 13). In these six sentences which JH and BW annotated as Previous work (1), KP chose Issues (2), Results (4) or Analysis (5); following a group discussion on our annotations, KP stated she would change them all to (1), which would have further improved the inter-annotator agreement to 93%. In one of these sentences (3-5), without a citation, BW found it difficult to assess whether the authors were discussing their current results, stating facts generally known in their field, or referring to a previous study (the sentences before and after did have citations): *Disassembly of actin fibers with cytochalasin D causes the accumulation of actin containing aggregates*. Ultimately BW annotated 3-5 as Previous (1). Although KP later stated she would change from Results (4) to Previous (1) for this sentence, JH in retrospect said he would change from (1) to

(4): "...this is likely one of their own results as they elaborate on them in the following sentences"; this is yet another indication that there is 'intra-annotator' variation over time, and that even the two annotators who share a background in science will have different perceptions of which argument category applies. JH's decision, however, points to another problem: the following sentence (3-6) cites *[13]* the current authors' own previous (2002) work; but the Results category (4) must contain reference to their current experiment, so Previous (1) seems the most appropriate choice for sentence 3-5. Aside from the above group of JB~K sentences, there were only two sentences where we disagreed: one (BK~J) between Results (4) and Analysis (5), and one (JK~B) between Previous (1) and Results (4) (similar to the above situation for 3-5).

Under Model 2 we had total agreement on only six of the 27 sentences, and three-way disagreement on three (see Table 14). BW thought *The cycling of syntaxins would make it possible to...* (sentence 2-6) referred to the future, and thus was Problem in context (6); KP saw it as a Claim (1); JH saw it as a Qualifier (5) because "it is an interpretation of results". Although JH's choice is legitimate, it should be noted that this type of sentence could certainly be a Claim – a "proposition based on interpretation of results". After reviewing, BW felt that this sentence could be either Claim (1) or Qualifier (5), but not Problem in context (6) (finding this error was another instance of intra-annotator variation). The verb tense is significant in *Kinetic studies indicated that...[13].* (sentence 3-3): although JH classed this as Grounds (2), the past tense along with the citation indicate they are referring to (their) previous work, and thus Grounds (2) is not appropriate. KP saw this as a Warrant (3) i.e., a Claim made in a previous paper, and BW

saw it as Backing (4) i.e., specific data from an outside source. As biomedical researchers tend to work on the same issues over many years, it can be difficult, especially for non-specialists, to decide whether authors are referring to their current experiment, or their previous work. We also all disagreed on the final sentence (4-6): *We used...in our studies to preserve...and therefore made them more visible than if a standard...had been used.* BW felt it was Problem in context (6), but later clarification regarding this category made her realize that this was an error. KP chose Grounds (2), noting that she was "unsure where to put their simple methodology under this Model." JH had Qualifier (5) as it "is a statement on the validity of the experimental design of the current study"; (5) seemed reasonable, but more on the basis of its explanation for a result.

In fact there was some inter-annotator variation in all six sentences of the final paragraph under Model 2. Whereas articles generally start with background material, here the authors end their paper with descriptions of previous studies which in some way support their own findings. Thus our variation in sentences 4-1 to 4-5 was entirely between categories Warrant (3) and Backing (4), both of which refer to external evidence. There were seven sentences each for the JK~B and BK~J splits, and four with JB~K (see Table 14). JH found seven Claims, KP five, and BW only one (see Table 17). Although variation between (3) and (4) is not critical, this amount of variation related to Claims is definitely problematic: identifying Claims is at the core of Model 2.

2.3.2.3 Feedback following annotations of T3-T5

After reviewing the statistics he had compiled JH felt that a lot of the inter-annotator variation under Model 1 stemmed from our "misunderstanding of the material" and the "ambiguity/overlap" in the Model categories. He observed that there was considerable confusion regarding whether a text referred to current Results (4) or Previous work (1), and between statements of Results (4) vs. Analysis of results (5). In his view, the use of verbs such as *suggest (that)* and *indicate (that)* (examples of hedges (Table 3)) did not necessarily imply that a sentence should be annotated as (5) rather than (4). JH felt strongly that we should have clearer, more "comprehensive" definitions of each category before any revisions to the Models were made.

KP compared our annotations under Model 2. Her recommendation was to create a more hierarchical version of the Model i.e., split categories into sub-categories. For example, there would be two distinct types of Claim: a Claim based on current results, and a Claim based on previous work, or different types of Problem in Context: those based on existing knowledge, and those focussed on future studies. Further, she stated her belief that Models 1 and 2 should be combined to create a hybrid Claims-based one (see Graves' response below).

In order to get feedback from Graves on whether we were applying her Model correctly, I sent her copies of T4 as annotated by myself and KP (JH's annotations were delayed). In response to KP's suggestion regarding combining the two Models, Graves wrote:

> I read [KP's] comments about combining the two Models, but the problem with that is that they are coding different parts of the argument. [Model 1] codes for general argumentative directions, while [Model 2] codes specific argumentative moves. You could combine them, but then some parts of the text would have to be coded twice: once identifying the type of argument and then again to identify the specific turn of the argument (for example, this statement is a claim for the authors' work [Model 2]; this claim is of the type that draws on the *topos* (topic) of "undisputed fact" [Model 1]). Do you see what I mean? They aren't dealing with equivalent aspects of the argument: the two Models do complement one another but they pay attention to different aspects of the argument.
>
> Personal communication, March 13, 2008.

At the level of specific annotations, she observed that BW was better at differentiating between Warrants (3) and Backing (4), but that KP was better at identifying Claims (1) the authors were presenting based on their own research: "I think this grows out of her more extensive experience with scientific prose (and your lack of background in the area)." She stressed that it is crucial to distinguish between a Claim made by authors, and the evidence on which it is based (Grounds). The problem is that a single sentence will not infrequently contain both: "The whole sentence may be the claim, but individual phrases in the claim are sometimes grounds (the evidence is stated as part of the claim)." This dilemma, and others like it, led me to develop a 'Trumping' system for dealing with argumentatively complex sentences in the final corpus (see Section 2.5 below). She also clarified for me that not every instance of Backing requires a specific Warrant.

2.4 Overview of phases I and II

2.4.1 Three-way inter-annotator Agreement

The striking fact that the average three-way agreement was virtually identical between phases I and II under both Models has already been mentioned in Section 2.3.2.2 (see Tables 6 and 12); in Table 18 below we see the average across the five training articles was 61.2% under Model 1 and 32.2% under Model 2. The rankings of individual articles within the corpus are also provided, with rank '1' having the highest total inter-annotator agreement. All three annotators found the content of article T2 to be the most accessible and understandable of the corpus, thus it is not surprising that it ranked highest under both Models. At the other extreme, we all found articles T1 and T3 to be the most difficult to understand and to annotate; they received the lowest rankings under Model 1 (fourth and fifth), while under Model 2 they were fourth and third respectively. The fact that T3 is the longest article, with 42 sentences, added to its difficulty. (Details on the problems encountered during annotations of T3-T5 have already been given in Sections 2.3.2.2.1-3.) Longer sentences are often more likely to be complex and therefore to have lower levels of inter-annotator agreement; article T2 has the longest average sentence length at 30 words, but under phase I sentence-splitting was allowed, thus the length cannot be viewed as a relevant factor. The shortest average sentence length of 23 words was in article T5; it ranked second for three-way agreement under Model 1 (although at 70% it is virtually tied with article T2 for first), but fifth under Model 2. Although sentence length may be a factor in particular cases, in general it is the overall level of difficulty of the article's content – technical jargon, unclear writing style, etc. – that is more important.

	BMC-series	Number of Sentences	Average Words/ Sentence	Total Agreement Model 1	Rank Model 1	Total Agreement Model 2	Rank Model 2
T1	**Molecular Biology**	31	25	52%	4	23%	4
T2	**Microbiology**	31	30	71%	1	42%	1
T3	**Chemical Biology**	42	27	50%	5	36%	3
T4	**Medical Genetics**	16	29	63%	3	38%	2
T5	**Cell Biology**	27	23	70%	2	22%	5
All		147	27	**61.2%**		**32.2%**	

Table 18: Total annotator agreement and rankings for Models 1 and 2 – training corpus

2.4.2 Hedges in the Training Corpus

Table 19 below presents all instances of hedging that were identified in the Discussion sections of articles T1-T5 (see Table 3 for the full preliminary list of hedges). Five of the six modal verbs were found (there were no instances of *should*), with *may* being by far the most common, with approximately half of all modals (18 of 35). They are highlighted in the shaded portion of the table. Six of the nine lexical verbs were represented, with *suggest* accounting for almost half of all lexical verbs (13 of 27). *Possible/possibly* and *likely* were the only hedging adjectives/adverbs found, and there were no hedging nouns. Verbs were overwhelmingly the preferred hedging choice, accounting for 79.5% of all hedges in the training corpus.

	T1	T2	T3	T4	T5	TOTAL
# sentences	31	31	42	16	27	147
May	8	3	1	5	1	**18**
Might	2	0	1	0	1	**4**
Could	3	0	0	1	0	**4**
Would	1	3	0	0	0	**4**
Can/cannot	1	2	1	1	0	**5**
Appear	1	0	0	0	2	**3**
Indicate	1	0	1	1	3	**6**
Perhaps	0	0	1	0	0	**1**
Seem	1	0	2	0	0	**3**
Suggest	2	4	1	0	6	**13**
Think	0	0	0	0	1	**1**
Possible/y	2	5	1	0	3	**11**
Likely	3	0	2	0	0	**5**
TOTAL	**25**	**17**	**11**	**8**	**17**	**78**

Table 19: Hedge distribution by article - training corpus

Not all articles in the corpus were equally 'hedged': the extremes are T1 with 25 hedges in 31 sentences (including eight instances of *may*) and T3 with only 11 hedges in 42 sentences. The Discussion section of T3 does contain a higher than usual percentage of straightforward statements of results, sentences which are less likely to be hedged. Nevertheless, it is not clear why there was such variation in hedging across the five articles i.e., how much was related to the novelty of the authors' work, the challenge their results presented to their field, personal writing style, etc. In Chapter 3 hedges are examined in more detail in the articles of the final corpus.

In reviewing the instances of *can* and *cannot* in the training corpus it became clear that not all were of the epistemic type (see Section 1.4.1.1). For example, in sentence 2-6 of article T1 the authors state: *However, strategies that exploit these differences, for example for developing new antibiotics targeted at a specific group of bacteria, can now*

be envisioned. Given the strength of the evidence in the previous five sentences, the meaning of *can* seems to be 'is now able to be envisioned' rather than 'may possibly be envisioned'. Given the polysemous nature of this verb i.e., it can express ability or permission as well as possibility, and the high probability of its occurrence as a non-hedge as in the example above, it was decided to eliminate it from the set of hedges for the final corpus.

2.4.3 Feedback on Argument Type

Both JH and KP felt that my three preliminary Argument Types (see Table 4, Section 1.5.2) were not appropriate for the articles in our corpus. In particular, as mentioned above, the notion of being 'novel' is a given, but it might be useful to develop sub-categories of 'novel'. Some examples discussed were: interpreting previous studies in a new way, using a new methodology, using previous results to design a new experiment. We also explored the possibility of differentiating 'revolutionary' (i.e., a paradigm shift) articles from 'evolutionary' ones, although we agreed that the latter are by far the more common. JH noted that in the training corpus he found authors attempting to "narrow the research question" or "eliminating possibilities". This type of refinement does exist in our corpus, but I felt that from the point of view of rhetoric, and given that we are limited to one Type per paper, it was most important to focus on the creative aspect i.e., what do these refinements signify to the field, or what have the authors done with this new information.

Following input from and discussions with JH and KP, and with the basis of rhetoric as the art of persuasion, I considered that Argument Type might be based on what I have found authors trying to 'sell' across *BMC* articles, rather than e.g., whether their results conflict with what previous researchers have found. Based on our training corpus as well as my previous exposure to *BMC*-series data I created the following non-exhaustive list of what authors are typically trying to market:

- Better utility
- More effective
- Wider applications
- Could lead to disease identification/treatment in humans
- Identify a significant new direction for future research
- Major implication(s) for future of the field

I believed that approaching argumentation from this point of view could give a more 'macro' description than the preliminary Argument Types in Table 4. Some aspects of the original Types are already captured at the 'micro' (sentence) level by categories such as Analysis (5) under Model 1 or Claim (1) under Model 2.

2.4.4 Problems with annotations

As referred to above, category (2) of Model 1 (Issues under dispute) did not seem to be a good fit for these data: it led to confusion with other categories, especially Previous work (1) and Results (4). My original thinking had been to separate out this category from

general background (with the focus on particular debates in the field which had motivated the current experiment), but what we generally found were statements of earlier work as it related to e.g., their choice of experimental technique (which could be categorized as Previous (1) or Methods (3)), or of their results in relation to e.g., expanding on or contradicting the work of other researchers in their field e.g., *we failed to replicate* (which could fit in Results (4)). I also realized that category (5) (Analysis) was problematic: In applying Model 2, where it became clear that a Claim could be based on the authors' current experiment, or on previous work, I saw that statements of 'analysis' were not necessarily restricted to their current results; in fact, presenting a new interpretation or understanding of previous work is an important aspect of supporting one's current argument. The inter-annotator variation between Results (4) and Analysis (5) could have been caused by being forced to choose one category where a single sentence contained material from both categories, by an annotator's understanding of verbs such as *suggest* (see 2.3.2.3 above), or simply by individual differences in interpretation of the material (which may be affected by one's view of the more 'macro' level of the argument; see below).

Variation under Model 2 was sometimes related to lack of familiarity with the Toulminian concepts or misunderstanding the categories, sometimes to not understanding the scientific content (e.g., BW failed to recognize a Claim (1) or took uncited statements to be Grounds (2) rather than Backing (4)), and sometimes, as with Model 1, to annotators' idiosyncratic understandings of either the Model or the data. One clear problem was the variation between Warrants (3) ("understanding of the problem based on

external evidence") and Backing (4) ("data and information from other studies"). Since they both refer to evidence external to the current study, for the purposes of this project it did not seem crucial to differentiate between these categories; as Graves put it, as long as we are agreeing on either (3) or (4), "that is close enough". We all agreed that it was difficult to categorize statements related to methodology (a specific category under Model 1), but the number of such sentences in our corpus was relatively small.

Through phases I and II of training, there were several questions raised concerning Model 2 that I was able to clarify, either with feedback from Graves, or from my own understanding as I continued annotating: category (6) (Problem in context) is always focussed on the future, and should not be used for existing debates in the field (see Sections 2.3.2.2.2/3 above). A Claim is a "proposition put forward based on analysis and interpretation of results" (my underlining for emphasis), not simply a statement of a finding in their study. For example, the first sentence of paragraph two in T5: *Our immunofluorescence microscopy studies indicate that engogenous syntaxin 2, 3 and 4 are located only in short sections of the plasma membrane and they are not dispersed all over the plasma membrane.* (2-1) was annotated as Grounds (2) by KP and BW, but JH felt it was a Claim (1). He defended his choice on the basis of the statement being "relevant to the title" and "a target of interest" for researchers. Both of these statements may be true, but they do not necessarily imply that a sentence is a Claim; this sentence describes findings presented in detail in their Results section, it does not involve analysis of these results, and thus is Grounds (2).

Although the distinction between Claim (1) and Grounds (2) became clearer, BW continued to have considerable difficulty trying to decide between the categories Claim (1) and Qualifier (5) for certain sentences. For example, sentence 1-4 from article T5 which follows two statements we all agreed were Grounds: *It is therefore possible that syntaxin 2 might cycle between the plasma membrane and the perinuclear membrane vesicles, and syntaxin 3 between the plasma membrane and the TGN in NRK cells.* (hedges are underlined): This would seem to be a Claim - a proposition, which may or may not be true, based on analysis of their current findings in 1-2 and 1-3. At the same time it seems to be a Qualifier - a "possible explanation for their data". In another example, BW thought the following sentence (3-8) from T5 was a Qualifier: *This suggests that syntaxin 2 and 3 are not as tightly attached to actin as syntaxin 4.*; she saw this as a possible explanation. The medical science students, however, both annotated the above two sentences as Claims. Thus BW's lack of background in a biomedical field could mean that she did not recognize the rhetorical significance of these types of statements. The other aspect of a Qualifier, to "compare and contrast with external evidence", led to difficulty with complex sentences: such a sentence could include current data (Grounds (2)) as well as external evidence (Warrants (3) or Backing (4)), but its main overall rhetorical purpose could be to compare the internal and external data, i.e., to act as a Qualifier (5). The annotator must decide which of these four categories would be the most appropriate.

There is an additional issue that exists when selecting argumentative roles at the sentence level: how is the choice of category affected by the surrounding text? This may concern

the 'macro' level of argument structure, such as that presented in the Model 2 flowchart in Table 2; such a 'big picture' seems especially crucial when attempting to correctly identify a Claim, where understanding the argument flow may help to decide whether a statement is a Claim or a Qualifier (see above). An article's discourse is also organized structurally, where each paragraph typically represents a theme or idea, and has its own flow. For example, the final paragraph of article T5 consists largely of a series of statements describing the results of a variety of previous studies; KP felt that although individual sentences seemed to qualify under Model 1 as Previous work (1), "the overall purpose of the paragraph was to present Issues under dispute". This is a dilemma not easily solved where the argument category is to be selected only at the sentence level.

An additional problem may arise when attempting to correctly identify cohesive argument structure at the discourse level. The scope of a sentence-initial demonstrative pronoun may be ambiguous. For example, a sentence beginning *This indicates that* may follow a series of sentences in a paragraph; it is not always clear whether *This* refers to the single preceding sentence, or to several. Thus from the argument perspective, it may not be obvious if the authors are discussing internal or external evidence, or both. Knowledge of the biomedical field may be critical in resolving these types of anaphora.

2.5 Revisions

In the following two Sections I describe the process of amending both preliminary Models of argument as presented in Tables 1 and 2; the revised Models as applied to the final corpus are found in Appendices C and D. Here I introduce a convention for both

Models 1 and 2: to avoid confusion with the preliminary Model categories, all revised Model category names will appear in small capital letters e.g., CONTEXT, CLAIM. It is hoped that this will also alleviate the problem of changes in numbering e.g., in the preliminary Model 2 the Qualifier category was number (5), whereas in the revised Model 2 the QUALIFIER category is number (4). Also new in the revised Models was the introduction of 'Trumping' guidelines. These were to be used where an annotator encountered a complex sentence, or when they were simply conflicted between two (or more) categories, i.e., category x Trumps category y. It was hoped that this would reduce the inter-annotator variation by controlling for some degree of subjective differences.

Given what I had learned during the training phases, there were two key questions to address in revising the Models of argument: are the categories clearly defined such that annotators can readily distinguish between them, and do the categories seem to be a good 'fit' for biomedical research texts. Annotators may still be conflicted between argument categories for a particular sentence, but this should not be because they are unclear regarding the specifications for these categories.

2.5.1 Model 1

A key insight I gained during the training process was the understanding that Model 1 is essentially information-based, and that what biomedical researchers need to be able to do is to separate 'old' from 'new' information. In my interviews with research scientists at UWO it had become clear that as specialists in a particular field, these researchers are already familiar with most of the background material in these articles; they want to be

kept up-to-date on new (and credible) work in their field. Hence, it was decided to create a new category called 'CONTEXT' which focuses on background material to the current study, including statements describing work done previously: in other words, 'old' information. Material that was in the original category (1) (Previous work) would now belong to this new CONTEXT (1) category.

As mentioned above the original category (2) (Issues under dispute) was problematic throughout the training annotations. It was difficult to distinguish it from Previous work (1) (such issues could be longstanding) or Analysis (5) (an analysis of current results could form part of a dispute in the field). Based on the notion of distinguishing new from old information, most text that would have been annotated as (2) under the original Model 1 would now belong to the CONTEXT (1) category i.e., it describes conflicts already existing at the time of the authors' study. The original category (5) (Analysis) was expanded so that it was not limited to the results of the current experiment; an understanding of Claims in Model 2, as well as observing how science is argued in academic writing, made it clear that authors are frequently presenting analyses not only of their current results, but of work done previously, or both. Thus the revised category (5) ANALYSIS covers analysis of current *or* earlier results, and therefore may encompass material that previously would have been seen as Issues under dispute (2). In addition, text was added (see Appendix C) to make it clear that statements focussing on the significance of their findings or any limitations of their study belong to the ANALYSIS category.

The original category (3) (Methods) became the revised category (2) METHOD, but remained essentially unchanged. It was also decided that the original category (4) (Results of current experiment) would be subdivided. Given that we now wanted to separate out texts providing only new information, the new category (3) (CURRENT RESULTS) was limited to results of the authors' current experiment, and statements that describe results of the current experiment specifically in relation to other studies would now belong to a new category (4) (RESULTS COMPARED).

The Trumping guidelines for Model 1 were generally based on the priority of information gain e.g., CURRENT RESULTS (3) would Trump CONTEXT (1). In addition, the ANALYSIS category (5) may Trump categories (3) (CURRENT RESULTS) or (4) (RESULTS COMPARED): Experience with the *BioMed*-series of articles has shown that generally the most crucial statements in the authors' argumentation – where they present what their results *mean*, and stress the originality of their ideas – are more often those that interpret their findings or suggest what these imply, rather than straightforward presentations of their results. Nevertheless, especially with complex sentences, there may be a situation where an annotator feels that the analysis aspect is weaker or less crucial than the results component, and hence may select (3) or (4) rather than ANALYSIS (5). All Trumping is at the annotator's discretion, and should be applied keeping in mind their understanding of the Model being applied as well as the authors' overall argumentation. The revised Model 1 includes the Trumping guidelines for annotators and is presented in Appendix C; the colours there are those employed by annotators to electronically highlight the corpus texts.

2.5.2 Model 2

Following our experiences with annotating the training corpus, I changed the format of Model 2 from a flowchart to a closed list of category choices (see Appendix D). Although the original Toulmin Model is based on the format of logic ('if p then q'), and the Graves adaptation (see Table 2) follows a directional flow in the structure of an argument, I wanted to make it clear that our approach was not of the 'decision tree' Model. Despite the fact that we are at some level taking into account the 'macro' level of argumentation, the annotation for each sentence is chosen by a linear search of the six available categories to decide which best applies (and using the Trumping guidelines where appropriate). This decision process is thus symmetric to that for Model 1, except it has only five categories available whereas Model 2 has six. Also I had decided that annotators would have to categorize all sentences in the final corpus, under both Models. (Note that the new EXTRANEOUS (0) category under Model 2 (see below) still qualifies as an annotation; during the training process, sentences had been initially allowed to remain unannotated.)

The major alteration to the preliminary version of Model 2 in Table 2 was the addition of a new category 'EXTRANEOUS' (0); during the training phases, and in consultation with Graves, we agreed that it is possible that some of the text in the Discussion section of our articles is extraneous to the authors' core argumentation. Although one could argue that text would not be included unless it had some argumentative purpose, we all felt that we had seen 'background' material that did not fit into the Toulminian structure of argument

(including that of Graves' adaptation). This would seem to be relatively common in this type of academic writing, particularly as a 'preamble' to their current study and its Claims; for example, statements of the authors' motivation for their particular study may be background, and not part of an argumentative move. Any sentence fitting this category was to be left with no colour highlighting in the final corpus. For statistical purposes this would be recorded as a category (0) to indicate it was not part of the numeric ((1)-(5)) representations of sentences within the authors' argumentation.

Category (1) (CLAIM), the most significant category in Model 2, was not fundamentally revised. Text was added to clarify that Claims could be based on work done previously as well as results of their current study (see Appendix D). Given that in our training corpus we felt some Claims were extremely important to the article's argumentation whereas others seemed less crucial, subcategories of 'Major' and 'Minor' Claims were added to the text of the revised Model 2; it was not necessary, however, to indicate this differentiation in annotating.

Category (2) (GROUNDS) remained unchanged. The original categories (3) (Warrants) and (4) (Backing) were merged to form the new category (3): WARRANT/BACKING. The original (3) described the "understanding of the problem based on external evidence", whereas (4) had been for the "external evidence" itself. This distinction led to some confusion, with ongoing difficulties in identifying the difference between the two categories, especially as statements often seemed to fit in either or both. Given this, and the fact that for our purposes the main goal was to differentiate between internal and

external evidence, we agreed to create the single revised category (3) (see Graves personal communication in 2.4.4 above).

The preliminary category (5) became category (4) (QUALIFIER) in the revised Model 2; it remained essentially unchanged. Graves' original text was maintained, with the addition of a clarification regarding how a QUALIFIER might function in relation to a CLAIM. Although there were difficulties during the training phase with this category, it was clear that we had to do better with understanding its role and applying it rather than attempting to alter its specifications. In the original version of Model 2 category (6) (Problem in context) was divided into two subsections: (6a) "Ways that the Claim qualifies or impacts the larger problem" and (6b) "New directions for additional research on the larger problem". After training annotations it came to light that JH and KP had not realized that (6a) and (6b) were to be differentiated; it was decided that in the revised Model this distinction did not need to be maintained. The text for the new PROBLEM IN CONTEXT (category (5)) includes the clarification that this category must relate to the authors' current results, and that it is forward-looking from their study.

The Trumping guidelines for Model 2 are based on the primacy of CLAIMS, thus category (1) Trumps all others. If a sentence seems to include external evidence as well as a comparison to their internal evidence, QUALIFIER (4) should Trump WARRANT/BACKING (3), and material from their current study should Trump previous work. The revised Model 2, including colour codes and guidelines for Trumping, is presented in Appendix D.

2.5.3 Hedges

The only change made in the list of hedges between the training and final corpora was the addition of the adverb *unlikely* (the negation of *likely*, already on the list in Table 3) and the removal of the verb *can* as discussed above in Section 2.4.2. Despite the fact that only 13 of the listed 30 hedges were found in the training corpus (and almost 80% of them verbs), the corpus was small (only five articles), and it seemed worthwhile to maintain a relatively large set of hedging possibilities for investigation in the final (and larger) corpus. Thus, the set of hedges sought in the final corpus was identical to the preliminary list in Table 3 with the exception of *can* and *unlikely*: five modals, nine lexical verbs, nine adjectives/adverbs and six nouns.

2.5.4 Argument Type

After considering my attempt to categorize arguments according to what researchers are trying to 'sell' (in 2.4.3 above), and as a result of further discussions with JH and KP during the training phase, it became clear to me that what I crucially wanted to capture in an article's Argument Type is the most significant aspect of its novelty. Toulmin identifies this core newness in the physical sciences as "discovery". Although written well before our twenty-first century ways of doing and writing science, and at a time when Toulmin wondered "how far are [genes and electrons] thought of as really existing, and how far as mere explanatory devices" (1953: 11)), his notion of 'discovering' still seems to apply to our *BMC*-series corpus articles. In the mid-twentieth century Toulmin stated: "The heart of all major discoveries in the physical sciences is the discovery of

novel methods of representation, and so of fresh techniques by which inferences can be drawn – and drawn in ways which fit the phenomena under investigation." (1953: 34) Today, however, scientists have access to very sophisticated tools which allow them to see, study and quantify entities that were unknown 50 years ago; methods of representation are now only one possible aspect of scientific novelty. In addition, discoveries are being made at a far faster rate, partly because information technology allows the rapid and wide-ranging dissemination of research results.

Even with this core focus on discovery, it is still possible to have a virtually infinite number of classes of novelty; based on my experiences with *BMC* articles, however, I identified four key classes. Below in Table 20 is the revised list of Argument Types available to annotators for the final corpus:

1) Advanced/improved methodology or experimental design

2) New creation/concept

3) New way of looking at/interpreting/evaluating existing data/previous results

4) Leads to/opens up new research direction, refines an existing research question, or contributes to addressing a significant research issue

Table 20: Revised list of Argument Types

Type (4) is more general than Types (1), (2) and (3); my expectation was that if none of the first three Types seemed to fit a given article, Type (4) would apply.

2.6 Other Issues

Although revisions were made to the Models, including the development of the 'Trumping' system, there remained issues over which I did not have control: most significant were those related to the corpus data and the annotators. As with any text created by someone other than the reader, it is impossible for the reader to get inside the writer's head; we can only create our personal understanding of what the text means to us. Trying to analyze or interpret how authors decided to structure their argumentation, from the point of view of either of the Models, is not easy, especially when one is not an expert in the particular biomedical field. If there is inter-annotator variation in assessment of the authors overall rhetorical strategy, it seems likely that this might lead to inter-annotator variation at the sentence level. In particular, under Model 2 an understanding of the argument structure of an article should help in identifying Claims.

In addition, as discussed above in 2.4.4, we must take into account the problems of context. What effect does the choice of a particular category for a sentence have on the preceding, or the following, sentence? For example, is a Claim (1) more likely to be followed by a Qualifier (4) than Grounds (2)? If a sentence has no citation, but the sentence(s) before and/or after do have, it can be impossible for the non-expert to know if current or previous results are being referenced. For example, in article T5 (sentence 3-9) BW found the term *Madin-Darby canine kidney epithelial cells* introduced. There was no citation for this sentence, but a search of the entire article found no other mention of this term; also, since the following sentence (*<u>These</u> vacuoles are associated with...[28]*) did

have a citation, she assumed that 3-9 referred to [28] i.e., previous/external work. However, both JH and KP saw 3-9 as describing current work: they annotated it as Results (4) (Model 1) and Grounds (2) (Model 2). It is not known how an expert in the field of Cell Biology (and syntaxin 2 and 3) would annotate this sentence, but there is clearly a difference between BW on the one hand, and JH and KP on the other.

Thus two crucial variables are: How comfortable is an annotator with understanding a) the concepts and categories of the Models of argument, and b) the background and technical content of *BMC*-series articles? BW was considerably more familiar with (a), but totally out of her depth with (b); JH and KP were new to (a), but had four years of exposure to scientific texts. Unfortunately this study did not include any true 'experts' in biomedicine as annotators to serve as a benchmark for understanding the technical content of our corpus data; our only comparisons are thus between a non-expert (BW) and two Medical Science students (JH and KP). There are also the unknowns one finds with annotators who work in isolation e.g., Did they follow instructions properly? Did they read the entire article, or at least ensure that they were covering all the material that was pertinent to the Discussion section? Did they spend sufficient time and care on the task? JH and KP's performance during the training phase led me to believe that they were sufficiently invested in the project to be able to produce results from the final corpus that would be worthy of detailed investigation.

CHAPTER 3 ANNOTATION OF FINAL CORPUS

3.0 Introduction

This Chapter presents the results of the annotations of the final corpus. First I look at inter-annotator agreement and variation across the twelve articles, and break down the number of sentences where we all agreed, under both Models, by argument category (Section 3.3.1). I then present the distribution of all annotations (not only those on which we had three-way inter-annotator agreement) under Models 1 and 2 by argument category, and by annotator (Section 3.3.2); these latter data provide a comparison of individual coder patterns. Following this in Section 3.3.3 are detailed discussions on each of the twelve articles where I provide example sentences which are representative of the types of disagreement encountered under each Model, as well as a brief description of any characteristics which make the article unusual e.g., very long sentences, densely hedged, etc. Then in Section 3.3.4 I report on the distribution of hedges found across the corpus, and break down the hedged sentences by Model and category, and annotator agreement groupings. Finally I present the distribution of Argument Types by article in the final corpus, and by annotator (Section 3.3.5).

3.1 Corpus

The final corpus is composed of twelve articles randomly selected from the BioMed Central database (www.biomedcentral.com). As was the case with the training corpus (see Section 2.1) all articles are from the *BMC*-series; the complete list of journal names and article titles (with URLs) for the final corpus is to be found in Appendix D. This

corpus covers nine different journals across a range of biomedical research fields; they are listed in Table 21 below. All conventions are the same as those presented in 2.1; the articles will be referred to by the codes C1 through C12 as noted in Appendix D. Only the Discussion sections of the articles were annotated (see Section 2.1.1).

Colour-annotated documents were crucial to me in my detailed analyses of the individual articles; spreadsheet data were useful for overall distributions and calculations, but visually comparing annotations of the same text, in context, was very useful. In addition, in-text citation numbers were purposefully left unannotated making it easy for me to identify them when I was reviewing the articles; as will be noted below in Section 3.3.3, citations are often an important factor in argument categorization. In order to give the reader an idea of what our annotated data look like I have included an example Discussion section (article C10) with colour annotations under both Models by the three annotators; these are found in Appendices G to L. The annotator's initials, the Model number and Argument Type are displayed above the Discussion section.

3.2 Instructions to Annotators

There were four steps in the annotations of the final corpus: 1) choose the article's Argument Type from the list in Table 20 (Section 2.5.3), 2) colour annotate the Discussion section under the revised Model 1 and Model 2 (as in Appendices D and E) making use of the Trumping guidelines where necessary, 3) enter the argument category codes for each sentence into an SPSS spreadsheet for each of the twelve articles, under both Models, and 4) enter each occurrence of a lexical hedge from the set in Table 3

(with the minor revisions in Section 2.5.3) into the SPSS sheet. More detailed instructions as they were provided to annotators are found in Appendix F.

The decision was made to use SPSS rather than e.g., Excel to record data for this study as it provides a wider range of statistical capabilities and is the 'industry standard' for research in the Social Sciences. Although this study is a pilot project testing the application of our two Models of argument through annotating a relatively small corpus, it is believed that the data collected in this annotation project could be used for statistical analyses in future projects beyond the scope of the current study. In addition all three annotators had worked with SPSS prior to this study.

3.3 Results of Annotations

3.3.1 Overview of inter-annotator agreement and variation

JH performed annotations on the final corpus during the final exam period and with a deadline of an out-of-country job shortly after exams finished; this suggests he was more rushed to complete his work than KP who remained in London for the summer term and submitted her annotations at a later date. On completion all documents were sent to BW electronically. Unlike during the training process, feedback from annotators was not part of the final corpus, thus in the balance of this Chapter only the results rather than the process of making annotation decisions are available (except for BW).

I gathered the data from the three annotators and compiled statistics on agreement and variation by annotator and by Model. The length (in number of sentences) of the Discussion sections ranged from 21 (C7) to 49 (C2) with a total of 400 sentences in the final corpus. The number of sentences within a Discussion section where there was overall (three-way) inter-annotator agreement under each Model were calculated as a percent of all sentences in the section. These were then ranked (within Model) with the article with the highest level of overall agreement being first. To see if sentence length might be related to inter-annotator agreement – shorter sentences may be less likely to be complex – the average number of words per sentence was calculated. These data are presented below in Table 21.

Measuring total inter-annotator agreement is an important metric in evaluating both Models of argument; high levels of agreement amongst three different annotators suggest that the categories are clearly defined and appropriate for the corpus data. Of course, there is no way of assuring that we are all 'right'; it may be that we are all 'wrong' in our choice, but simply agree in our error. Nevertheless, given the experience of training with applying both Models, the assumption is that if we all agree, we are likely to have made an appropriate choice.

	BMC Series	Number of Sentences	Average Words/ Sentence	Total Agreement Model 1	Rank Model 1	Total Agreement Model 2	Rank Model 2
C1	Biochemistry	30	29	80%	2*	23%	8
C2	Biochemistry	49	23	53%	7	51%	3
C3	Plant Biology	25	31	44%	10	20%	10
C4	Chemical Biology	24	24	58%	6	8%	11
C5	Plant Biology	33	24	36%	11	33%	7
C6	Physiology	45	22	47%	9	69%	1
C7	Physiology	21	19	81%	1	62%	2
C8	Neuroscience	35	27	80%	2*	34%	6
C9	Cell Biology	36	26	69%	3	39%	5
C10	Medical Genetics	27	21	52%	8	44%	4*
C11	Infectious Diseases	41	23	68%	4	44%	4*
C12	Molecular Biology	34	23	65%	5	21%	9
Avg		**33.3**	**24**	**60.5%**		**39.3%**	

Table 21: Total annotator agreement and rankings for Models 1 and 2 by article (* indicates a tie)

It was of interest in this study to look at inter-annotator variation as well as agreement: by identifying the sources of variation it was hoped that in the future variation could be reduced, and thus that total inter-annotator agreement could be increased. As will be discussed in detail in Chapters 4 and 5, the results of this study show two types of variation: random and systematic. The latter includes variation stemming from, for example, argument categories whose specifications overlap, causing confusion for annotators. In the balance of this Chapter I present data on variation, broken down by the factors which influence these results e.g., problematic argument categories, annotator

bias, and corpus article. Human-annotated corpora can only be useful as training data for IE systems if the systematic inter-annotator variation is reduced sufficiently; note that there is continuing debate in CL about how to calculate and evaluate the appropriate levels of agreement for different tasks (see Section 5.1 for a detailed discussion on this topic). Below in Tables 22 and 23 the same conventions for identifying two-way inter-annotator agreement groups are used as in the training corpus (see Section 2.3.1.3).

Given that the average overall inter-annotator agreement was higher for Model 1 at 60.5% (242 out of 400 sentences) than Model 2 at 39.25%[3] (157 of 400) there was 21.25% more variation under Model 2. All annotators had found Model 2 more difficult to apply than Model 1 during training, and with revisions, Model 2 still had six categories (EXTRANEOUS (0) was a legitimate choice) and Model 1 only five, thus there were more opportunities for variation under Model 2. Since average overall agreement varied, there was considerable difference in the number of sentences with two-way variation: under Model 1 there were 143 and under Model 2 almost half again as many at 210. Although the three sub-groups of two-way variation were almost exactly equally distributed under Model 2, they were skewed toward BK~J under Model 1 where this group accounted for almost half (69) of the 143 sentences. We had total (three-way) inter-annotator variation on only 3.75% of the 400 sentences under Model 1 and 8.25% under Model 2. These data are presented below in Tables 22 and 23 for Models 1 and 2 respectively:

[3] The total percent in Table 21 is rounded to 39.3%; here I include the data to two decimal places.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	TOTAL	
All agree	24	26	11	14	12	21	17	28	25	14	28	22	**242**	**60.50%**
All disagree	0	7	1	0	3	1	0	0	3	0	0	0	**15**	**3.75%**
JK~B	1	3	7	7	4	0	0	0	2	3	1	4	**32**	**8.00%**
JB~K	2	5	3	1	1	8	1	2	3	4	7	5	**42**	**10.50%**
BK~J	3	8	3	2	13	15	3	5	3	6	5	3	**69**	**17.25%**
TOTAL	**30**	**49**	**25**	**24**	**33**	**45**	**21**	**35**	**36**	**27**	**41**	**34**	**400**	**100.0%**

Table 22: Number of sentences in agreement groups by article – Model 1

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	TOTAL	
All agree	7	25	5	2	11	31	13	12	14	12	18	7	**157**	**39.25%**
All disagree	6	0	3	9	1	1	0	4	1	1	2	5	**33**	**8.25%**
JK~B	2	13	12	2	5	2	1	10	11	7	1	5	**71**	**17.75%**
JB~K	9	6	1	8	6	4	4	4	6	2	7	11	**68**	**17.00%**
BK~J	6	5	4	3	10	7	3	5	4	5	13	6	**71**	**17.75%**
TOTAL	**30**	**49**	**25**	**24**	**33**	**45**	**21**	**35**	**36**	**27**	**41**	**34**	**400**	**100.0%**

Table 23: Number of sentences in agreement groups by article – Model 2

Given that I was interested in comparing the performance of the two Models, it was important to break down the sentences on which we all agreed by argument category: Do some categories seem easier to agree on than others? Do certain categories thus make one Model easier to apply than the other? Of course these statistics must also be considered in light of a) the fact that there is less three-way agreement under Model 2 and b) the overall distribution of all sentences (not only those on which we all agree) by category. (The

latter are presented in Tables 26 and 29 below in Section 3.3.2.) Recall that all revised Model category names are presented in small capital letters, as in Appendices C and D.

Under Model 1 the category on which we most frequently agreed was ANALYSIS (5) with 108 of the 242 sentences (44.6%); next was CONTEXT (1) at 24.8% and CURRENT RESULTS (3) at 17.8%. METHOD (2) accounted for 8.7% and RESULTS COMPARED (4) only 4.1%. The category on which we agreed most often under Model 2 was GROUNDS (2) with 43 of the 157 sentences (27.4%). The next four categories were relatively evenly distributed: EXTRANEOUS (0) at 18.5%, WARRANT/BACKING (3) at 17.8%, CLAIM (1) at 15.3% and QUALIFIER at 14.6%. PROBLEM IN CONTEXT (5) accounted for only 6.4 % of the total agreement sentences. These data are presented below in Tables 24 and 25 for Models 1 and 2 respectively:

	CONTEXT (1)	METHOD (2)	CURRENT RESULTS (3)	RESULTS COMPARED (4)	ANALYSIS (5)	TOTAL
C1	10	1	3	1	9	**24**
C2	5	6	4	0	11	**26**
C3	4	1	2	0	4	**11**
C4	4	0	0	2	8	**14**
C5	5	0	2	1	4	**12**
C6	4	0	4	0	13	**21**
C7	3	1	6	1	6	**17**
C8	4	3	7	0	14	**28**
C9	0	2	3	2	18	**25**
C10	5	0	4	2	3	**14**
C11	13	2	4	0	9	**28**
C12	3	5	4	1	9	**22**
TOTAL	**60**	**21**	**43**	**10**	**108**	**242**
Percent	**24.8%**	**8.7%**	**17.8%**	**4.1%**	**44.6%**	**100%**

Table 24: Number of sentences with total annotator agreement by category – Model 1

	EXTRANEOUS (0)	CLAIM (1)	GROUNDS (2)	WARRANT/ BACKING (3)	QUALIFIER (4)	PROBLEM IN CONTEXT (5)	TOTAL
C1	0	4	2	1	0	0	**7**
C2	9	3	6	6	1	0	**25**
C3	1	2	1	0	0	1	**5**
C4	0	1	0	0	0	1	**2**
C5	6	2	1	0	1	1	**11**
C6	2	5	11	10	1	2	**31**
C7	1	2	6	0	3	1	**13**
C8	0	1	6	1	3	1	**12**
C9	0	3	5	1	4	1	**14**
C10	4	0	3	1	3	1	**12**
C11	3	1	0	8	6	0	**18**
C12	3	0	2	0	1	1	**7**
TOTAL	**29**	**24**	**43**	**28**	**23**	**10**	**157**
Percent	**18.5%**	**15.3%**	**27.4%**	**17.8%**	**14.6%**	**6.4%**	**100%**

Table 25: Number of sentences with total annotator agreement by category – Model 2

3.3.2 All annotations by argument category

Each of the three annotators selected a single argument category for each of the 400 sentences in the final corpus, thus there is a total of 1200 annotation tokens (3 x 400) for each of Models 1 and 2. Thus in Tables 26 and 29 below the "TOTAL" column for each of the twelve articles in the corpus will contain a value three times the number of sentences in Table 21. The distribution of categories varies among articles depending on the writers' argumentative structure and goals, writing style, field, etc. Some of these particularities will be discussed in Section 3.3.3 below where detailed analyses of the individual articles are given, but here I will present the overall category distributions by article and by annotator for both Models 1 and 2.

3.3.2.1 Model 1

As shown in Table 26 below, the most frequently selected category across the corpus under Model 1 was ANALYSIS (5) with 36.0% of all tokens (432 of 1200), followed by CONTEXT (1) at 28.0% (337 of 1200). This is not surprising given that typically the Discussion section's goals are to situate the authors' work in their broader field and, more crucially, to provide interpretation and analysis of their results. The next most frequent category was CURRENT RESULTS (3) with 189 tokens (15.8%); there was a wide range among articles on this category with C4 having no annotations for category (3) (see Section 3.3.3.4) and C7 having 30% of its 63 tokens as (3). The next category was METHOD (2) with 10.7% of all tokens (128 of 1200); this also showed a considerable range from C4 which had no category (2) tokens to C12 which had 22.5% of its 102 tokens as (2). The least represented category was RESULTS COMPARED (4) with 9.5% of all 1200 annotations. This latter is to be expected as comparisons to previous results are frequently found in complex sentences accompanied by some form of analysis or speculation; in these cases ANALYSIS (5) would generally Trump (4) (but see below). The full distribution of annotation categories by article under Model 1 is displayed below in Table 26:

	CONTEXT (1)	METHOD (2)	CURRENT RESULTS (3)	RESULTS COMPARED (4)	ANALYSIS (5)	TOTAL
C1	36	3	10	6	35	90
C2	42	24	19	11	51	147
C3	22	7	16	10	20	75
C4	20	0	0	10	42	72
C5	44	15	9	9	22	99
C6	28	19	30	14	44	135
C7	14	3	19	6	21	63
C8	18	11	23	4	49	105
C9	10	10	12	13	63	108
C10	31	2	18	16	14	81
C11	54	11	15	5	38	123
C12	18	23	18	10	33	102
Total	**337**	**128**	**189**	**114**	**432**	**1200**
Percent	**28.0%**	**10.7%**	**15.8%**	**9.5%**	**36.0%**	**100%**

Table 26: All annotations by category and article – Model 1

Although in Section 3.3.1 above one can see statistics on argument categories where all annotators agreed under Model 1 (Table 24) and on variation by annotator grouping (Table 22), it was also of interest to see if particular annotators were more inclined to choose some categories more than others. In fact the data show a surprising – given that we had some type of variation on 39.5% of the sentences under Model 1 – degree of similarity on how often we identified particular categories across the corpus. This distribution by annotator is presented below in Table 27:

	CONTEXT (1)	METHOD (2)	CURRENT RESULTS (3)	RESULTS COMPARED (4)	ANALYSIS (5)	Total
BW	121	39	59	36	145	400
JH	92	43	67	57	141	400
KP	124	46	63	21	146	400
TOTAL	**337**	**128**	**189**	**114**	**432**	**1200**

Table 27: Total category frequencies by annotator – Model 1

All three annotators selected category (5) (ANALYSIS) virtually the same number of times and the range of variation for categories (2) (METHOD) and (3) (CURRENT RESULTS) was very small. The only notable difference was JH's choosing CONTEXT (1) for approximately 30 fewer sentences than BW and KP, with the offset of his annotating approximately 30 more sentences as RESULTS COMPARED (4); these cells are shaded. Note that the data in Table 27 only show the number of times each annotator chose a category across the corpus; they provide no evidence as to where annotators agreed on category for particular sentences. This latter information showing pair-wise inter-annotator category choices across the final corpus for Model 1 is found in Section 4.2.1.3.1 in Tables 51-53.

It was also of interest to see what relation exists between the sentence categories on which all annotators agreed under Model 1 (Table 24) and the overall category choices presented in Table 26 above. In order to compare these two sets of data I use percentages: with total inter-annotator agreement in Table 24 there is a total of 400 (sentences) whereas for all annotations in Table 26 there are 1200 tokens (c.f. Section 3.3.2 above). This comparison is presented below in Table 28:

	CONTEXT (1)	METHOD (2)	CURRENT RESULTS (3)	RESULTS COMPARED (4)	ANALYSIS (5)
Entire corpus	28.0%	10.7%	15.8%	9.5%	36.0%
Total agreement	24.8%	8.7%	17.8%	4.1%	44.6%

Table 28: Category frequencies for all sentences and those with total annotator agreement - Model 1

Given that under Model 1 total annotator agreement occurred in 242 sentences or 60.5% of the corpus, one would expect some similarity in the frequencies of categories between the two groups noted above; in the two rows of Table 28 above one sees both similarities and differences. It is not surprising that we agreed on fewer of the category (4) sentences (RESULTS COMPARED) as these tend to be complex sentences, often including a CURRENT RESULTS (3) or ANALYSIS (5) component. In addition, the Trumping guidelines state that ANALYSIS (5) could Trump (3) or (4) "if the sentence is critical to the argumentation" (although there may have been doubt about how to identify 'critical' from 'non-critical' statements). Although there were overall annotations of 36% for the ANALYSIS (5) category, 44.6% of the sentences we all agreed on were in this category. It was my experience from discussions with annotators during the training period that they found this category more clearly defined than some others, and I believe that revising the ANALYSIS category to encompass previous work made sentences of this type easier to agree on. This pattern is also reflected in the pair-wise annotator comparisons (Tables 51-53 in 4.2.1.3.1) where in each of the three cases the ANALYSIS (5) category accounted for the largest number of sentences with pair-wise agreement.

3.3.2.2 Model 2

As shown in Table 29 below, under Model 2 the categories QUALIFIER (4) and EXTRANEOUS (0) were the most frequent annotations overall with 21.3% and 20.8% respectively. It is noteworthy both that greater than one-fifth of all sentences of our

Discussion sections were considered to not be part of the authors' argumentation, and that this category was close to being the most commonly selected of all six categories in Model 2. The next most common categories GROUNDS (2) and WARRANT/BACKING (3) had almost identical percentages in the corpus: 18.2 % and 18.0% respectively. This suggests that internal and external evidence appeared equally in support of the authors' CLAIMS. Category (2) GROUNDS exhibits a wide range of variation among articles from 0% in C4 to 30% in C6 and C7. CLAIMS (1) accounted for 15.4% of the 1200 tokens, ranging from 2.5% in article C10 to 27% in C9. PROBLEM IN CONTEXT (5) was by far the least commonly selected category at 6.3% in the corpus. Of its total of 76 tokens, 19 were identified in article C12, an unusually high number; if these were removed from the calculation, the percentage for category (5) would fall to 4.8%. Overall, except for PROBLEM IN CONTEXT, one sees a relatively equal frequency under Model 2 of categories (0) through (4), varying over a range of only 5.9%. This is in marked contrast to Model 1 where the category frequency varies from 9.5% to 36.0% for a range of 26.5% (see Table 26). The distribution of argument categories by article under Model 2 is presented below in Table 29:

	EXTRANEOUS (0)	CLAIM (1)	GROUNDS (2)	WARRANT/ BACKING (3)	QUALIFIER (4)	PROBLEM IN CONTEXT (5)	TOTAL
C1	24	20	11	19	15	1	**90**
C2	42	24	32	29	18	2	**147**
C3	25	10	12	13	7	8	**75**
C4	15	12	0	14	19	12	**72**
C5	33	11	7	18	20	10	**99**
C6	7	30	41	35	16	6	**135**
C7	12	8	19	3	16	5	**63**
C8	13	23	29	11	26	3	**105**
C9	10	29	21	13	32	3	**108**
C10	19	2	18	18	21	3	**81**
C11	24	8	12	36	39	4	**123**
C12	26	8	16	6	27	19	**102**
Total	**250**	**185**	**218**	**215**	**256**	**76**	**1200**
Percent	**20.8%**	**15.4%**	**18.2%**	**18.0%**	**21.3%**	**6.3%**	**100%**

Table 29: All annotations by category and article – Model 2

The frequency of category choices by annotator under Model 2 (Table 30 below) show more inter-annotator dissimilarities than under Model 1 (Table 27). There is striking variation in the number of sentences annotated as EXTRANEOUS (0): JH identified 116 sentences as (0), KP 80 and BW only 54; as a partial balance to this, JH had only 49 sentences as WARRANT/BACKING (3), more than 30 fewer than either KP or BW. The question of whether statements of external evidence do or do not form part of the authors' argumentation is a difficult one that may involve a) knowledge of the field, b) an understanding of the Toulminian notion of argument structure and c) subjectivity. It is not clear whether BW's finding fewer EXTRANEOUS (0) sentences is based on better familiarity with argumentation or inadequate knowledge of the scientific fields, or both. It is worth noting, however, that even though JH and KP share a similar academic background in medical sciences, JH found almost half again as many EXTRANEOUS

sentences as KP. Category (0) under Model 2 is classically a case where determining the one 'correct' annotation seems extremely difficult, if not impossible.

The other noteworthy variation under Model 2 is in the categories of CLAIM (1) and QUALIFIER (4): KP identified 91 CLAIMS, almost as many as JH and BW combined; on the other hand, she only had 50 sentences annotated as QUALIFIER, close to half the number that JH and BW identified. The fact that JH and BW, who do not share an academic background in biomedical sciences, had such similar numbers in these categories as compared to KP suggests that factors other than scientific knowledge are playing a large part in this variation. The issue of variation between CLAIM and QUALIFIER will be discussed further in Section 4.2.1.2. The number of category choices across the corpus by annotators under Model 2 is presented below in Table 30, with the notable values discussed above shaded for emphasis. In parallel with Model 1, note that the data in Table 30 only show the number of times each annotator chose a category across the corpus; they provide no evidence as to where annotators agreed on category for particular sentences. This latter information showing pair-wise inter-annotator category choices across the final corpus for Model 2 is found in Section 4.2.1.3.2 in Tables 54-56.

	EXTRANEOUS (0)	CLAIM (1)	GROUNDS (2)	WARRANT/ BACKING (3)	QUALIFIER (4)	PROBLEM IN CONTEXT (5)	Total
BW	54	45	86	81	108	26	400
JH	116	49	61	49	98	27	400
KP	80	91	71	85	50	23	400
TOTAL	**250**	**185**	**218**	**215**	**256**	**76**	**1200**

Table 30: Total category frequencies by annotator – Model 2

In Table 31 below I break down the number of CLAIMS identified by each annotator by corpus article. Although some of these data will be mentioned in discussions of individual articles in Section 3.3.3 below (e.g., JH annotating no CLAIMS in C10 and C12, the range from 4 to 14 in number of CLAIMS in C9) here I want to draw attention to the wide variation found between articles, as well as by annotator, in the number of CLAIMS, ranging from a total of 2 in article C10 to 30 in C6. Although one could expect some of this range to be accounted for by the length of the Discussion sections (they range from 21 to 49 sentences) it is evident in Table 31 that this is not the only factor. In order to compare across articles I have calculated the total number of CLAIMS for all annotators ("TOTAL") as a percent of all possible tokens for each article ("Total tokens" = 3 X the number of sentences). From this perspective C10 has the lowest percentage of CLAIMS at only 2.5% of all possible annotations, and C9 has the highest at 26.9%. As shown in Table 29 as well, overall CLAIMS accounted for 15.4% of the total possible annotation tokens in the final corpus.

	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**	**C12**	**Total**
Total tokens	90	147	75	72	99	135	63	105	108	81	123	102	1200
BW	4	4	4	4	2	10	2	5	4	1	2	3	**45**
JH	7	7	2	1	4	7	2	6	11	0	2	0	**49**
KP	9	13	4	7	5	13	4	12	14	1	4	5	**91**
Total	**20**	**24**	**10**	**12**	**11**	**30**	**8**	**23**	**29**	**2**	**8**	**8**	**185**
Percent	22.2	16.3	13.3	16.6	11.1	22.2	12.7	21.9	26.9	2.5	6.5	7.8	15.4

Table 31: Number of CLAIMS identified per article by each annotator - Model 2

In parallel with Model 1, and using the methodology in Section 3.3.2.1 above, I now compare the percentages of sentences where we all agreed under Model 2 (Table 25) with the overall categories for all annotators in Table 29. These data are below in Table 32:

	EXTRANEOUS (0)	CLAIM (1)	GROUNDS (2)	WARRANT/ BACKING (3)	QUALIFIER (4)	PROBLEM IN CONTEXT (5)
Entire Corpus	20.8%	15.4%	18.2%	17.9%	21.3%	6.3%
Total Agreement	18.5%	15.3%	27.4%	17.8%	14.6%	6.4%

Table 32: Category frequencies for all sentences and those with total annotator agreement - Model 2

Despite the fact that overall inter-annotator agreement is lower under Model 2 (at 39.3%) than Model 1, there are similarities in the frequency of some categories between the agreed-upon sentences and the overall corpus annotations. As shown above, the percentages for category EXTRANEOUS (0) are very similar and those for categories CLAIM (1), WARRANT/BACKING (3) and PROBLEM IN CONTEXT (5) are virtually identical. It is interesting to note that this similarity in percentages for CLAIM exists despite the extreme inter-annotator variation on CLAIM shown in Table 31 above. The similar percentages for EXTRANEOUS (0) also mask the fact that this category was involved in some inter-annotator variation in 101 sentences in the final corpus (see Table 48 Section 4.2.1.2). The GROUNDS (2) and QUALIFIER (4) percentages are inversely distributed with total agreement on 9.2% more GROUNDS (2) sentences and 6.7% fewer QUALIFIER (4) sentences. The QUALIFIER (4) category was a source of considerable inter-annotator disagreement in the final corpus (see Section 4.2.1.2 and Table 49).

3.3.3 Results by Article

Below are more detailed descriptions of the results for each of the twelve articles in the final corpus. They contain general data regarding an article's particular structural and/or rhetorical characteristics and its situation relative to others in the corpus e.g., its Model 1 agreement percentage is similar to that for Model 2. In addition, example sentences are given which are representative of the types of inter-annotator variation and agreement found across the corpus under each of the Models. For reference purposes note that for all articles the number of sentences, average sentence length, three way agreement percentages and rankings for both Models are found in Table 21 above.

3.3.3.1 Article C1 *BMC Biochemistry*

The Discussion section of C1 contains 30 sentences with an average length of 29 words per sentence, longer than average. The authors present a forceful rhetoric using the phrase *strongly suggest* on four occasions. There is considerable referencing of previous work, especially in paragraph four, with a total of 17 citations in the Discussion section. Two of the sentences that caused inter-annotator variation under both Models are of interest. Sentence 1-3 cites a previous work from 1994, one of whose authors is also one of the current (2007) authors: *It is noteworthy that E1 could come from both sources, the local transformation of...or from...[20].* My underlining is to emphasize that the present tense is followed by the modal *could*, thus compounding the complexity raised by the citation: is the focus here more on background, or their current experiment (in sentence 1-2 they discuss the *local conversion* in reference to their current results)? The following sentence

It is well recognized that... cites two additional papers, then sentence 1-5 states *Cells that possess steroid sulfatase could use...to produce E2.* Since there is no citation in 1-5, it was unclear to BW, who was unfamiliar with the scientific terminology, whether 1-5 is expressing a new speculation, or a stating an existing possibility presented in the works cited in 1-4.

Under Model 1 the total annotator (three-way) agreement was extremely high at 80%, tying it with article C8 for rank of second-highest agreement under Model 1 (Table 21). There were no sentences where we all disagreed, and the two-way agreement sentences were fairly evenly distributed: three BK~J, two JB~K and one JK~B (see Table 22). The main variation, as in sentences 1-3 and 1-5 above, was between CONTEXT (1) and ANALYSIS (5).

There was total agreement on only seven of the 30 sentences (23%) under Model 2, ranking it at eighth (see Table 21). This striking difference with Model 1 at 80% is largely explained by the type of sentences, such as 1-3 and 1-5 described above, as well as those in paragraph four (see above), and how they relate to the different Models of argumentation. In the former group, the variation under Model 1 was typically between categories CONTEXT (1) and ANALYSIS (5) (see above), whereas under Model 2, any one of four categories might apply: EXTRANEOUS (0) (referencing previous results, not part of the argument), CLAIM (1) (a proposition based on previous results), WARRANT/BACKING (3) (evidence from an external source) OR QUALIFIER (4) (a possible explanation, or a comparison between internal and external evidence). The six sentences with three-way

disagreement under Model 2 were all of this type. Of the ten sentences in paragraph four, only the tenth had total agreement under Model 2; one had three-way disagreement, and eight had two-way agreement where BW and JH classed them as EXTRANEOUS (0) but KP saw them as WARRANT/BACKING (3). These accounted for eight of the nine JB~K sentences (see Table 23). Note that under Model 1 these eight sentences were classified by all annotators as CONTEXT (1).

3.3.3.2 Article C2 *BMC Biochemistry*

The Discussion section of article C2 has the largest number of sentences – 49 – with an average length of 23 words per sentence. The Results section (which precedes the Discussion) is long (7 pages) and contains numerous graphs and images, thus there are many aspects of analyses presented in the Discussion. BW found the scientific content almost impenetrable, and thus struggled with annotating it, especially under Model 2. The total annotator agreement percentages were almost identical – 53% for Model 1 and 51 % for Model 2 – but the rankings were quite different: seventh and third respectively (see Table 21).

Under Model 1 there were seven sentences (14%) with total disagreement, all of which involved at least one annotator choosing CONTEXT (1) where others saw the sentences as belonging to another category. Two sentences provide examples of this three-way disagreement. Sentence 5-2 follows a sentence describing the methodology of their current experiment (*...with a variety of single point mutants...*): *Because each of the single point mutants has a differing enzymatic activity and propensity to interact with*

associated protein, the combined data was expected to reveal the factors governing trypsin sensitivities. This sentence provides uncited factual information about the *mutants*, but also a statement of hypothesis going into their current experiment (*was expected to*). Despite the Trumping guide of (2) through (5) Trumping (1) (see Appendix C), BW felt strongly that the overriding characteristic of 5-2 was of 'old information', and thus annotated it as CONTEXT (1). Based on the speculative component, JH's choice of ANALYSIS (5) seems reasonable, as does KP's choice of METHOD (2); thus, all three annotation choices seem legitimate. Sentence 6-3 contains a similar statement of hypothesis: *Consequently, if associations with...dictate trypsin sensibility, the expectation was that the wild type and single point mutants would display significantly different sensitivities to trypsin.* Our annotations here had the same three-way split as 5-2. These are classic examples of the type of content complexity where total annotator agreement is difficult to achieve under Model 1; under Model 2, however, we all agreed that these sentences were EXTRANEOUS (0) i.e., not part of the authors' current argumentation. The largest two-way split under Model 1 was BK~J with eight sentences (see Table 22), six of which involved BW and KP identifying the sentence as CONTEXT (1) whereas JH chose either RESULTS COMPARED (4) or ANALYSIS (5).

Under Model 2 there was overall agreement on 25 of the 49 sentences, and no cases of overall disagreement. The two-way agreement was skewed toward JK~B with thirteen of the 24 sentences (see Table 23); of these, six JH and KP annotated as EXTRANEOUS (0) whereas BW believed them to be part of the argumentation, and JH and KP agreed on CLAIM (1) for four sentences which BW had annotated as QUALIFIER (4). This skewing

may be due in part to JH and KP sharing a greater understanding of the material being discussed, which BW found extremely inaccessible (as noted above). For example, BW annotated sentence 6-1 *The experiments with the single point mutants are also helpful in illuminating the role of protein association in governing trypsin digestion susceptibilities* as GROUNDS (2), internal evidence; JH and KP, however, categorized 6-1 as EXTRANEOUS (0), not part of the argument structure. There were four sentences we all agreed were METHOD (2) under Model 1, but under Model 2 we had only two-way agreement: BW had all as GROUNDS (2) but JH and KP had EXTRANEOUS (0). One such sentence (4-2) exemplifies the difficulty in dealing with statements of methodology under Model 2: *Since HDAC1 phosphorylation promotes HDAC1 enzymatic activity and protein associations [24], we studied HDAC1 mutants lacking phosphorylation sites.* This can be seen as simply background for their choice of method (EXTRANEOUS), but by referencing a previous study, it can also be seen as using external evidence to justify this choice i.e., it is part of the authors' argumentative strategy, where either GROUNDS (2) or WARRANT/BACKING (3) could apply. There was a wide range of variation in the number of sentences identified as CLAIMS (1): KP classified thirteen sentences as (1), JH had seven CLAIMS, and BW had only four (see Table 31).

3.3.3.3 Article C3 *BMC Plant Biology*

C3 has the highest average sentence length in the final corpus: 31 words per sentence; there are 25 sentences in the Discussion section, close to the average of 24. It is singular in the corpus in having subheadings to structure the text, at the beginning of paragraphs two, three and six. It is unusual in two other respects: paragraph five has only one

sentence (with 35 words), and sentence 6-2 is extremely long – 5.5 lines and 48 words. Given that longer sentences are more likely to be grammatically and/or argumentatively complex, which in turn leads to a greater likelihood of inter-annotator variation, one could predict poor inter-annotator agreement for this article; if fact, this was the case. Under Model 1 all annotators agreed on only 44% of sentences and under Model 2 it was only 20%; they shared ranking of tenth, the second worst agreement figures (see Table 21). I note that these rankings are in spite of the fact that for BW at least the content was relatively accessible: she understood concepts such as *aphid resistance* and *dwarf seedlings*.

Although the three-way agreement was poor under Model 1 at 44%, we had three-way disagreement on only one sentence (4-2): *Resistance-breaking biotypes of A. idaei have been recorded on 'Autumn Bliss' which carries* A_{10} *but not on 'M. Leo' which carries both* $A_1 A_{10}$*; this could be a consequence of gene pyramiding in the latter. Autumn Bliss* and *M. Leo* are not part of their current experiment, and despite the *have been recorded* there is no citation (in fact there are no citations in the five sentences of paragraph 4); the final clause of 4-2 presents speculation by the current authors. All agreed that sentence 4-1 was CONTEXT (1), and for BW sentence 4-2 was essentially background to the possible future directions presented in the next three sentences, i.e., it was also CONTEXT (1). KP's annotation of ANALYSIS (5) for 4-2 also seems valid, but JH's choice of RESULTS COMPARED (4) seems inappropriate as there is no reference to their current results. The very long sentence (6-2) compares their current results (*consistent with*) to two other previous *models*; this is a classic case of RESULTS COMPARED (4) (as annotated by BW

and JH). KP's choice of CURRENT RESULTS (3) for 6-2 misses the rhetorically crucial issue of positioning their results in the wider field, but the Trumping guidelines state that (3) should Trump (4), especially if the new results are not presented elsewhere. This suggests a weakness in the 'new information' focus of Model 1: if (3) Trumps (4), then important external evidence may be missed. The largest two-way split under Model 1 was JK~B with seven sentences (see Table 22): these covered a range of differences, but most commonly JH and KP chose CONTEXT (1) where BW's annotation was CURRENT RESULTS (3) or RESULTS COMPARED (4).

Under Model 2 we had total agreement on only five (of 25) sentences and total disagreement on three (see Table 23). Two of the latter group are discussed in the paragraph above. Despite the lack of citation BW annotated 4-2 as evidence external to the study (WARRANT/BACKING (3)); although JH had not chosen CONTEXT (1) under Model 1 (see above), here he annotated 4-2 as EXTRANEOUS (0); KP believed it was a statement of CLAIM (1). As discussed above, sentence 6-2 would seem to fit squarely into category (4) QUALIFIER under Model 2 ("compare and contrast with external evidence"); this was BW's annotation. JH, however, despite choosing RESULTS COMPARED (4) under Model 1, here categorized 6-2 as WARRANT/BACKING (3), thus missing the argumentatively crucial aspect of comparison to what the current authors had *observed.* KP saw this sentence as a CLAIM (1); however, what the authors *observed...indicates that just one gene segregated in this progeny, consistent with...* seems more a straightforward statement of what they saw rather than a "proposition based on analysis of results". Close to half the sentences under Model 2 (12 of 25) involved the two-way split JK~B (see

Table 23): seven cases JH and KP saw as EXTRANEOUS (0) when BW saw them as part of the argumentation, and JH and KP annotated sentences 2-3, 2-4 and 2-5 as GROUNDS (2) where BW believed they were QUALIFIERS (4).

3.3.3.4 Article C4 *BMC Chemical Biology*

The Discussion section of article C4 is relatively short, 24 sentences with an average length of 24 words. It is unique within the final corpus for two reasons: Firstly, it is the most heavily hedged article, with 21 hedges, including 11 instances of the modal verb *may* (see Table 33 below); this seems to be based on the fact that their results are *unexpected* (3-1) and challenging to the more standard models (*alternative to current proposals* (4-1)). The authors set the hedging tone in their opening sentence: *Barring the discovery of a novel, unsequenced or unidentified protein or peptide, these data point to the possible sequence of LMWCr fractions and may point to new strategies in therapeutic design*; although the items I have underlined are not all in the list of hedges in Table 3, they all represent a deference to their community and a distancing from expressing categorical certainty about their results. Secondly, the Discussion section does not contain straightforward statements of their results; under Model 1 no one annotated a sentence as either METHOD (2) or CURRENT RESULTS (3) and under Model 2 there were no annotations as GROUNDS (2) (see Tables 26 and 29 above). In fact their title mentions only their *search*, not their results; this could be seen as another form of hedging. C4 is also unusual in that the fourth (penultimate) paragraph contains a series of four (4-3 through 4-6) 'rhetorical questions': The authors state that their new model *points directly back to significant...questions* (4-2), the implication being that these have not yet been adequately answered; the four questions are then posed to the reader.

Under Model 1 there was total agreement on 14 of the 24 sentences (58%), giving it a rank of sixth out of eleven. There were no sentences with three-way disagreement and the largest two-way split was JK~B with seven sentences (see Table 22). Included in this category were sentences 4-3 to 4-6 discussed above: BW saw them as CONTEXT (1) to their experiment, but JH and KP annotated them as ANALYSIS (5); in retrospect BW feels that as "speculation", ANALYSIS (5) might be the more apt category. BW's switch to ANALYSIS (5) on these sentences would improve the overall agreement to 75%.

Under Model 2 there was overall inter-annotator agreement on only two sentences, the first and last of the Discussion section; at 8% this article has the lowest ranking in the final corpus, and the worst performance across both corpora under either Model (see Tables 18 and 21). As described above, this article is unusual in a number of respects, at least some of which seemed to emphasize some of the difficulties with Model 2 which were encountered in other articles. For example, since the focus was clearly not on the results of the current experiment, but rather on situating their work in the context of other work in the field – there are 20 citations across paragraphs two and three – there was variation regarding whether text was part of the argument or not: BW and JH annotated sentences 1-2 through 2-4 as EXTRANEOUS (0) but KP saw 2-1 to 2-4 as WARRANT/BACKING (3) and 1-2 as PROBLEM IN CONTEXT (5). (These five sentences were part of the eight with the JB~K split, the majority of the thirteen sentences with two-way agreement in Table 23). In addition, the category GROUNDS (2)is one of the easier to agree upon, and there were no sentences in this category (see above). Of the nine

sentences with total disagreement (the highest number in the corpus under either Model), five were in the problematic block (see above) of 4-2 through 4-6: all were annotated as WARRANT/BACKING (3) ("understanding of the problem based on external evidence") by BW, QUALIFIER (4) by JH and as PROBLEM IN CONTEXT (5) by KP. No category seems to clearly fit these sentences, and one could argue for any of WARRANT/BACKING (3), QUALIFIER (4) or PROBLEM IN CONTEXT (5) as being appropriate.

Another major source of variation under Model 2 was found in the identification of CLAIMS (1): KP annotated seven sentences as CLAIMS, BW four, and JH only one (see Table 31). In the majority of these cases, the variation was between (1) and (4) (QUALIFIER). One example was sentence 4-8: *Alternatively, Cr(III) clusters may be transported non-specifically in serum by proteins, possibly including transferrin and serum albumin.* Although this particular (1) vs. (4) variation existed across the corpus, it seems to be more pronounced in C4 because of the high number of hedges, especially the verb *may*: many of these sentences could readily be seen as "possible explanations for their data" (QUALIFIER), or as a "proposition put forward based on analysis and interpretation of results" (CLAIM). This issue will be discussed in more detail later in Section 4.2.1.2.

3.3.3.5 Article C5 *BMC Plant Biology*

This is the only article in the training and final corpora that is a 'Methodology' rather than 'Research' article (the latter seem to be far more common in the *BMC*-series of journals). It has an unusual text structure: only two, very long, paragraphs, the first with

16 sentences and the second with 17; this seems to be more a matter of the authors' writing style than its being in the 'Methodology' category. The average sentence length is 24 words, the corpus average (see Table 21). Their results here are clearly an extension of work in which the various authors have been involved: of the 47 works cited in the entire article, 13 include one or more of the current authors, from 2003 to 2006, and one as yet unpublished (C5 was published in 2007). The Discussion contains 20 citations, and the focus is on positioning their results within the context of previous work, both their own and others'. BW found the technical content relatively accessible, and her difficulties in annotating were not related to problems in understanding the science.

Under Model 1 the three-way inter-annotator agreement was only 36% (12 of the 33 sentences), giving it the lowest ranking in the final corpus, and the worst performance across both corpora under Model 1 (see Tables 18 and 21). The primary source of inter-annotator variation was between CONTEXT (1) and other categories, especially METHOD (2) and RESULTS COMPARED (4). This problem is largely a reflection of the nature of the content as discussed above i.e., most of the text refers to external sources; overall, annotators selected CONTEXT (1) 44% of the time and CURRENT RESULTS (3) only 9% (see Table 26 above). Sentences 1-1 through 1-10 provide background to their study and the methods used, some with citations. For example sentence 1-9: *In Arabidopsis it is possible to reliably score embryo lethals, a phenotype resulting from deficiency at any of hundreds of genes*. For BW this was a statement of background fact i.e., CONTEXT (1), and KP agreed; JH, however, annotated it as METHOD (2). In fact for all of the first ten sentences BW had CONTEXT (1) and JH had METHOD (2); since (2) includes a "basis for

choice" of the researchers' current methodology, it seems that either category could apply for this block of sentences. KP annotated six of these as CONTEXT (1), three as METHOD (2) and one (1-7) as CURRENT RESULTS (3). This latter was an error. Although 1-7 refers to some of the current authors (*We*), the verbs are in the past tense: *We experienced such a difficulty with rice, where multiple attempts resulted in...(...unpublished results)*; these are previous, not current, results. In addition to the six sentences in the first paragraph there were seven additional BK~J sentences for a total of thirteen, by far the largest two-way agreement category (see Table 22). BW and KP annotated three of these (2-14, 2-15 and 2-17) as ANALYSIS (5) where JH chose CONTEXT (1). As with the variation above between CONTEXT (1) and METHOD (2), it is not straightforward to decide which category (CONTEXT (1) or ANALYSIS (5)) is more appropriate. For example sentence 2-14 evolves from discussions of previous work in 2-11 and 2-12: *An increasing number of concurrent mutations are of course less and less likely to be caused by the mutagenic treatment*; although ANALYSIS (5) seems reasonable ("suggest why something did happen"), the use of the present tense could also imply that this is more a statement of fact (*of course*), and thus CONTEXT (1).

The overall agreement at 33% under Model 2 was similar to that for Model 1, but the ranking was better at number seven (see Table 21). Eight of the 21 two-way agreement sentences under Model 2 had variation between EXTRANEOUS (0) and another category. Although JH had annotated 1-1 through 1-10 as METHOD (2) under Model 1, under Model 2 he categorized them all as EXTRANEOUS (0); BW annotated 1-3 and 1-4 as WARRANT/BACKING (3), and the balance as EXTRANEOUS (0); KP chose QUALIFIER (4) for

1-5 and GROUNDS (2) for 1-7 (reflecting the error described above for Model 1), and EXTRANEOUS (0) for the balance of 1-1 to 1-10. Under Model 2 the main block of two-way agreement was in sentences 2-5 through 2-12 and involved variation mainly between the categories CLAIM (1) and QUALIFIER (4), or between WARRANT/BACKING (3) and QUALIFIER (4); this variation stemmed at least in part from the particular characteristics of the content, as discussed above. As under Model 1 the largest two-way split was BK~J, with ten sentences (see Table 23). Sentence 2-13 was the only one where we all disagreed under Model 2: *For example, based on the Poisson distribution, the probability of obtaining two mutations in the same individual of either EMS- or Az-MNU-treated populations after screening 10 genes x 1,300 bp of DNA is 0.95*. BW found that these statistics seemed to come from their current study so annotated it as GROUNDS (2), JH saw it as a QUALIFIER (4) and KP felt it was EXTRANEOUS (0); again, the combination of lack of citation, the technical content, and their citing previous work in 2-10 and 2-11 leads to uncertainty regarding the most appropriate categorization.

3.3.3.6 Article C6 *BMC Physiology*

The Discussion section of C6 is the second longest in the final corpus with 45 sentences; they average 22 words per sentence i.e., are relatively short. The authors' results here are clearly an extension of work they had done in their lab in 2002, 2004 and 2005 (C6 was published in 2006) and they cite these previous works seven times in this section. Despite the fact that this continuity might imply certainty regarding the acceptance of their results, this article contains numerous (28) instances of hedging, including two sentences with three hedges (see Table 33 below). For example sentence 4-7 contains three from

the list of hedges in Table 3 (underlined for identification): *It is also possible that this lower MW was a degradation product of GIRK1, but we think this is unlikely because the same protein samples were used for determining GIRK2 and GIRK4 protein expression, and these samples showed no differences in MW.* They use the verb *believe* on five occasions, a hedge of which there was only one other instance in the entire final corpus. They also make use of a verb that I have seen so rarely, if at all, in my years of exposure to BioMed articles, that I never considered adding it to the list of commonly occurring hedges: They preface their speculation in sentence 2-6 with *We feel that...* Since the other paper (C7) in the corpus from the *BMC Physiology* series contains no instances of *feel* or *believe*, and has a lower frequency of hedging (see Table 33), it seems the findings above for C6 are related to writing style and/or specific content rather than being in a field with particular hedging requirements.

The overall inter-annotator agreement was quite poor under Model 1 at 47%, giving it a rank of ninth (see Table 21). Since BW found the technical content difficult, the expectation was that the JK~B agreement might be high, as the two science-oriented annotators would share a greater understanding of the material; surprisingly there were no sentences in this category. In fact the largest two-way split was in the BK~J agreement group with one third (15) of all sentences in the Discussion section. In this group the most common occurrences were where BW and KP annotated sentences as CONTEXT (1) but JH categorized them as METHOD (2), CURRENT RESULTS (3) or RESULTS COMPARED (4). Five such sentences occurred in the first paragraph where the authors provide background to their current experiment, much of which cites their own previous work.

For example, BW and KP chose CONTEXT (1) and JH chose RESULTS COMPARED (4) for sentence 1-8: *We previously first reported GIRK1 protein was seen in the three small cell lung cancer (SCLC) cell lines that express GIRK1 mRNA, and determined GIRK1 protein was not expressed in non-SCLC cell lines [10].* This is clearly a result of their 2005 study (*[10]*), with verbs in the past tense, and is presented before any specific results of their current experiment are mentioned; thus JH's choice of category (4) which compares current with past results seems inappropriate. There was only one sentence on which we all disagreed, 3-3: *In SCLC cell lines we saw expression only at 62 kDa [10].* Unlike sentence 1-8 above, however, 3-3 follows two sentences reporting on their current experiment (which we all agreed were CURRENT RESULTS (3)), where they state that one of the molecular weights they found was *62 kDa*; thus, BW annotated it as RESULTS COMPARED (4). This is a classic case of how the surrounding (here, previous) text affects the annotation: if one looks at the sentence in isolation, then it is similar to 1-8 above; in fact KP annotated it as CONTEXT (1), which in isolation could be correct. JH's choice of CURRENT RESULTS (3), however, is incorrect.

Under Model 2 the total inter-annotator agreement for C6 was 69%, the highest across both corpora, and ranking it first in the final corpus (see Tables 18 and 21). It is also unique across both corpora in that the Model 2 overall agreement is higher than that under Model 1 which was 47%. One evident reason for this is seen in the first paragraph: Under Model 2 category WARRANT/BACKING (3) includes "understanding of the problem" as well as specifics of external evidence; unlike the variation described above between CONTEXT (1) and other categories under Model 1, under Model 2 we could all agree that

these sentences were WARRANT/BACKING (3). Under Model 2 we had three-way disagreement on only one sentence, 2-4: *Since the predominant GIRK heterotetramers seem to be GIRK1/2 and GIRK 1/4 [reviewed in [11]], we concentrated on GIRK1, GIRK2, and GIRK4 protein expression in these cells.* The previous sentences discuss earlier work, with four different citations, leading to this statement regarding a reason for the methodology of their current experiment. BW annotated it as GROUNDS (2), JH as WARRANT/BACKING (3) and KP as EXTRANEOUS (0); the latter seems inappropriate given that justifying where they focused their study is a significant part of their argumentation. There were a high number of CLAIMS (1) annotated in C6: BW identified ten, JH seven and KP thirteen (Table 31). The most striking result, however, was that we had three-way agreement on five CLAIMS, the largest number in the corpus (see Table 25), which contributed to the high level of overall agreement for C6.

3.3.3.7 Article C7 *BMC Physiology*

Article C7 has the shortest Discussion section in the final corpus with 21 sentences. It also has the shortest average sentence length across both corpora at 19 words (Tables 18 and 21). It nevertheless includes the longest sentence in the corpus (2-6) at 66 words (not including the two citations) and nine tensed verbs. At the other extreme is sentence 1-7 with only five words: *Worker bumblebees are not sterile.* If sentence 2-6 is removed from the calculation as an outlier, the average sentence length becomes only 17 words. A smaller number of sentences yields fewer opportunities for inter-annotator variation and shorter sentences are less likely to be complex, also implying lower inter-annotator variation. In fact this was the case for article C7: Under Model 1 the three-way inter-

annotator agreement was 81%, the highest across both corpora, and ranking it first in the final corpus (see Tables 18 and 21). Although overall agreement under Model 2 was less at 62%, it ranked second highest in the final corpus (see Table 21). In addition C7 was the only article where there were no sentences with three-way disagreement under either Models 1 or 2 (see Tables 22 and 23). It should be noted that another key reason for the high agreement statistics was the material under discussion: there is very limited technical jargon, and BW found the content the most accessible of all 17 articles in the training and final corpora.

Given the high overall agreement there were only four sentences with any inter-annotator variation under Model 1 (Table 22). Sentence 1-10 follows sentences contextualizing their current result *...we started to see male eggs produced.* (1-6): *This egg production task would increase the variation in the days after injection factor, but would have no effect on our injection type result.* My underlining emphasizes that 1-10 clearly is discussing results from their current experiment i.e., *This* refers to 1-6, and follows a similar sentence (1-9) that we all agreed was ANALYSIS (5); BW and KP annotated 1-10 as ANALYSIS (5) as well, but JH annotated it as CONTEXT (1). Especially given that there is no citation here, it seems impossible to justify JH's categorization. Sentence 2-4 was another source of inter-annotator variation under Model 1. It follows sentence 2-3 where the authors cite previous results from Schmid-Hempel: *This does not explain the difference between our result and Schmid-Hempel's, as recently Freitak et al. found an increase in basal metabolic rate in a butterfly pupa due to encapsulation of a foreign particle [18].* BW and KP both annotated this sentence as RESULTS COMPARED (4) while

JH saw it as ANALYSIS (5); given that sentence 2-4 is both comparing current and previous results, and seeking an explanation for this divergence, either RESULTS COMPARED (4) or ANALYSIS (5) seem reasonable annotations.

Under Model 2 there were eight sentences with two-way split agreement, four where JH and BW agreed (JB~K), three BK~J and only one sentence where JH and KP agreed (JK~B) (see Table 23). Sentence 1-10 had disagreement parallel to that found under Model 1 above: BW and KP annotated this sentence as QUALIFIER (4) – speculation regarding the cause of a current result – but JH saw it as EXTRANEOUS (0); as above, this sentence is not 'extraneous', but rather forms part of the authors' line of argumentation, leading toward what we all agreed was a CLAIM (1) in sentence 1-12. The availability of the category QUALIFIER (4) under Model 2 led to three-way agreement on sentence 2-4: since QUALIFIER (4) includes aspects of both speculation by authors and comparison to external evidence, we did not have the type of variation found under Model 1 between RESULTS COMPARED (4) and ANALYSIS (5). As in C2 and C6 above, there was inter-annotator variation under Model 2 on a sentence that discusses the authors' current methodology: *Our study used LPS stimulation, which would lead to the production of antimicrobial peptides produced by the Imd pathway [17].* (2-3) Although this sentence cites a previous work, the focus is clearly on an aspect of the method used in their study. Under Model 1 all annotators agreed that 2-3 was METHOD (2); under Model 2 JH and KP annotated this sentence as EXTRANEOUS (0) but BW believed it to be a QUALIFIER (4) as it presents an explanation for a finding. It seems one could also argue for labeling this sentence as GROUNDS (2) based on their discussing *Our study*. In any case, these types of

sentences can be handled in a more straightforward manner under Model 1 with the availability of the METHOD category.

3.3.3.8 Article C8 *BMC Neuroscience*

The Discussion section of C8 has 35 sentences with an average length of 27 words (Table 21). It is the second most densely hedged article in the final corpus with 23 hedge occurrences, an average of one hedge every 1.5 sentences vs. the average of one every 2.1 sentences (see Section 4.3.1). These include eight instances of *may* and six of *would* (see Table 33 below). The authors' study involved the use of human fetal stem cells, a controversial methodology; it is not known whether the high frequency of hedging relates to this fact, or more generally to the presentation of results in the field of Neuroscience (this is the sole article from this *BMC*-series). The argumentation consists primarily of their presenting different possible explanations for their results, with the final paragraph (number 6) containing six of the total of eleven citations in the Discussion. The theme of exploring multiple possibilities may also be a factor in the high number of hedges.

Under Model 1 the total inter-annotator agreement was very high at 80%, tying it for the rank of second with C1 (see Table 21), and there were no sentences where we all disagreed. Of the seven sentences with two-way agreement two were in the JB~K category and five were BK~J. Despite the relatively technical content, there were no sentences where JH and KP agreed, and disagreed with BW (see Table 22). Given the authors' presentation as discussed above, the most frequent categories were CURRENT RESULTS (3) and ANALYSIS (5): paragraph two was universally annotated as CURRENT

RESULTS (3) and we all agreed that sentences 4-3 through 5-9 were ANALYSIS (5). The final sentence of the article (6-8) was, however, a source of variation: *The presence of v-myc cannot, however, be the primary determining factor because Cho et al. did not observe action potentials in v-myc derived human NSCs [7].* This is a classically complex sentence combining a previous result (*[7]*) with the ruling out of a possible explanation for their current results. BW and KP annotated it as ANALYSIS (5), (with BW conflicted between ANALYSIS and RESULTS COMPARED (4), but selecting ANALYSIS (5) based on the Trumping guidelines); JH, however, annotated it as RESULTS COMPARED (4). It is worth noting that if an annotator was not uncertain or conflicted regarding a sentence's categorization, they would not necessarily consult the Trumping guidelines; this may have been the case here for JH, or he may have considered that 6-8 was not "critical to their argumentation". Sentence 1-1 tells the reader from which areas of the brain the authors had selected the fetal stem cell lines; then: *These two brain areas were chosen as they are of particular interest as research models of cellular processes occurring during the development of degenerative diseases such as Huntington's, Alzheimer's and Parkinson's diseases.* (1-2) BW and KP annotated 1-2 as METHOD (2) ("basis for choice of methodology") but JH saw it as CONTEXT (1), possibly seeing it as "motivation for the current experiment".

The three-way inter-annotator agreement under Model 2 was only 34%, less than half of the 80% under Model 1, ranking C8 sixth in the final corpus (Table 21). The fact that this article presents considerable external evidence as well as a large component of speculation (see above) led to more variation under Model 2 as it was often not clear

whether a sentence discussing previous work was part of the authors' argumentation (was it EXTRANEOUS (0) or not), or if a statement was a CLAIM (1) or a QUALIFIER (4) ("possible explanation"). There were eight sentences with inter–annotator variation between EXTRANEOUS (0) and another category, and nine sentences with variation between CLAIM (1) and QUALIFIER (4). In eight sentences of the latter group there was total agreement on ANALYSIS (5) under Model 1. Although there were no sentences in the two-way JK~B group under Model 1, under Model 2 there were ten sentences of this type (see Tables 22 and 23), five of them where JH and KP annotated as CLAIM (1) and BW had QUALIFIER (4). Overall BW identified five CLAIMS, JH six and KP twelve (Table 31).

There was two-way agreement on sentence 1-2 under Model 2: In parallel with Model 1 above JH here chose EXTRANEOUS (0), but KP who had METHOD under Model 1 here had EXTRANEOUS (0), and BW annotated 1-2 as GROUNDS (2). Although under Model 1 this is a straightforward reason for their methodology (see above), under the Model 2 approach to argumentation the fact that the authors' study could ultimately help to understand a number of the most devastating human diseases is a crucial piece of support for CLAIMS they will make later in the Discussion, and thus a significant part of the argumentation, i.e., not EXTRANEOUS (0). The next sentence led to three-way inter-annotator variation: *Our initial hypothesis that morphological differentiation of the stem cells into neurons would also be reflected in their electrophysiological characteristics was not supported by the data.* (1-3) This provides a pre-experimental prediction as well as a general statement of result: BW annotated it as GROUNDS (2), KP saw it as a CLAIM (1) and JH as EXTRANEOUS (0). We also had total disagreement on Sentence 6-8 (see above): KP

annotated it as GROUNDS (2), JH as QUALIFIER (4) and BW believed it to be a CLAIM (1). Although sentences 1-3 and 6-8 were problematic under both Models, they caused two-way inter-annotator variation under Model 1, but total disagreement under Model 2.

3.3.3.9 Article C9 *BMC Cell Biology*

This article presents material that was essentially incomprehensible to BW, despite repeated readings; the Results section was long with many graphics, which were not helpful to her in understanding the Discussion. Even the title – Degradation of the LDL receptors by PCSK9 is not mediated by a secreted protein acted upon by PCSK9 extracellularly – which is also the first sentence of the Conclusion section of the Abstract, was confusing. It is stated as a negation (*not mediated*) but the scope was not clear to BW: But it is mediated intracellularly? (The final sentence of the Abstract states: *Rather, the PCSK9-mediated degradation of the LDLR appears to take place intracellularly...*; perhaps this is too hedged to be part of the title.) The fact that the authors are from a laboratory in Norway may also play a part i.e., they may not have English as a first language, or it may simply be their style of writing. In any case, the technical terminology was clearly a significant stumbling block for BW in trying to understand the researchers' argumentation.

The Discussion section of C9 is 36 sentences long with an average length of 26 words. There are few straightforward statements of their current results; instead there are numerous statements of speculation regarding the interpretations of these results, as well as citing previous works for direct comparison with their findings. As expected given this

type of presentation there is considerable hedging: 24 instances, or one hedge per 1.5 sentences, giving it virtually the same density as article C8 above (especially since their average sentence length is also virtually the same, see Table 21). *May* and *indicate* are the most common hedges in C9 (Table 33). The authors end the Discussion with a double hedge, followed by a standard researcher's caveat: *Thus it may seem that...*(5-7); *However, more studies are needed to determine the exact mechanisms by which...*(5-8).

Despite BW's difficulties with the content and writing style, the three-way inter-annotator agreement was relatively high under Model 1 at 69% (25 of 36 sentences), ranking it at third highest (Table 21). By far the most common category on which we all agreed was ANALYSIS (5) with 18 sentences; this was not surprising given the focus of the authors' presentation/argumentation as described above (see Table 24). There were numerous sentences of the 'current result implies interpretation' complexity, but it seemed that the guideline of ANALYSIS (5) Trumping CURRENT RESULT (3) where the sentence "is critical to their argumentation" was useful as there were only two sentences with inter-annotator variation between CURRENT RESULT (3) and ANALYSIS (5) (2-9 and 3-4). For example sentence 3-1 contains a current result followed by an implication of that result: *The failure of the secreted, truncated LDLR (EC-LDLR-His), which lacks the cytoplasmic and membrane-spanning domains, to be degraded when incubated with conditioned medium, indicates that PCSK9 does not degrade the LDLR directly by acting on the extracellular part of the LDLR*; we all annotated this as (5) rather than (3). One of the three-way disagreement sentences under Model 1 was 5-4: *This notion is supported by the findings of Benjannet et al. [4] who have found present PCSK9 in both early and*

late endosomes. This notion refers to a hypothesis presented in 5-3 (*seems to be involved in)*; BW annotated this as ANALYSIS (5) as supporting the speculation in 5-3, JH as RESULTS COMPARED (4), linking it to current results included in 5-3, and KP as CONTEXT (1). This type of sentence is difficult to categorize under Model 1, and is much better captured under Model 2 as external evidence important to their argument; in fact under Model 2 we all agreed on the annotation of WARRANT/BACKING (3).

Under Model 2 our overall inter-annotator agreement was only 39%, considerably lower than under Model 1, and ranking it fifth (Table 21). Although we all disagreed on only one sentence (1-2) we had two-way disagreement on 21 sentences, ten of which involved variation between CLAIM (1) and QUALIFIER (4). This latter on-going source of variation was particularly pronounced in C9 as there was such a wide difference in number of CLAIMS identified: BW annotated four sentences as (1), JH eleven and KP fourteen (see Table 31). Of the eleven sentences in the JK~B agreement group (Table 23) seven involved JH and KP agreeing on CLAIM (1) where BW selected another category. For example sentence 2-6 follows current results in 2-5 (which we all annotated as GROUNDS (2)): *This finding shows that PCSK9 purified by gel filtration degrades the LDLR when added back to cultured cells.* (2-6) BW believed this to be GROUNDS (2) ("internal evidence") but JH and KP annotated it as CLAIM (1) ("proposition put forward"). It seemed to BW that the authors were simply expanding on their results, but JH and KP appear to have viewed it as using internal evidence as a basis for a generalization ("based on analysis and interpretation of results"). This may be a case where BW's lack of knowledge of the scientific field meant that she misunderstood the statement. It remains

an open question, however, how often BW was not recognizing CLAIMS and/or how often JH or KP were over-identifying them. The first five sentences all involved variation between the annotation of EXTRANEOUS (0) and either GROUNDS (2) or WARRANT/BACKING (3). In two of these cases variation was avoided under Model 1 by the availability of the METHOD (2) category; for example we all agreed on METHOD (2) for sentence 2-2 under Model 1: *To answer this question, the effect of conditioned medium on the LDLR was studied after D374Y-PCSK9-His had been removed by affinity chromatography.* The prevalence of inter-annotator variation between EXTRANEOUS (0) and other categories under Model 2 suggests that the three annotators do not share the same understanding of argument structure.

3.3.3.10 Article C10 *BMC Medical Genetics*

The Discussion section of C10 has 27 sentences with an average length of 21 words – the second shortest across both corpora (see Tables 18 and 21); unlike in article C7, however, shorter sentence length has not led to high three-way annotator agreement (see Table 21). This relatively poor performance is also in spite of the fact that the content was at least partially accessible for BW as it discussed human subjects as well as genetics. Thus the inter-annotator variation may stem from lack of clarity in the annotators' shared understanding of the argument categories and/or difficulty mapping the particular content and writing style of C10 onto our Models of argument.

It is interesting to note that although the title mentions *stroke risk* first, and then *relationship with lipid profile*, there is no mention of *stroke risk* in the Discussion section.

In the Conclusion section of the abstract they state: *Our study does not support a major role for the ABCA1 gene as a risk factor for ischaemic stroke. Some haplotypes may confer a minor amount of increased risk or protection.* The first sentence does address the 'risk' issue, but hedges somewhat (*not...a major role*); the second sentence, however, is hedged almost to the point of vagueness. Although only the verb *may* is part of our Table 3 hedge list, all the items which I have underlined serve to anticipate possible challenges to their proposition; the possibilities of *risk or protection* seem to cover most, if not all, eventualities.

In the body of the article, however, they avoid the concept of 'risk' and simply address the association found; the Conclusion section begins: *In conclusion ABCA1 was not associated with ischaemic stroke in our population.* This rhetorical choice could be based on the apparent challenge their results present i.e., they have not confirmed what others have found. For example after citing previous work they state: *By contrast, our control group...* (1-6), *This has not been confirmed in our study, although...* (2-2) and *a protective role...has not been confirmed by all studies* (3-1). Only eight instances of the lexical hedges in Table 3 were found in the Discussion (see Table 33), but the authors are at pains to hedge their results by other means such as stressing the limitations of their study: *thus the results should be interpreted with caution* (4-4), *our study may not have been large enough to detect this* (5-2) and *possible confounders include* (5-6). We are given some further insight in their closing sentence into how the authors present their argument: *The changes in lipids post stroke remain controversial...* (5-8). This apparent controversy may be what is behind their reluctance to make CLAIMS of any sort; BW and

KP each identified only one CLAIM, and JH none, by far the lowest number in the final corpus (see Table 31). The issue of whether an article can have no CLAIMS will be addressed later in Section 4.2.1.3.3.

Overall inter-annotator agreement under Model 1 was 52%, ranking it at number eight (Table 21). There were no sentences where we all disagreed and the two-way agreement sentences were relatively evenly distributed: six in the BK~J category, four in JB~K and three in JK~B (see Table 22). Nine of the thirteen sentences with inter-annotator variation involved variation between CONTEXT (1) and other categories. For example, sentence 1-4 provides background material: *Other works have reported that CAD patients who are carriers of R219K allele had less severe arthrosclerosis [31] and overall lower risk of CAD [36].* BW and KP annotated this as CONTEXT (1) but JH believed it to be RESULTS COMPARED (4). His choice may have come from misreading *less* and *lower* as comparing to the current results; as there are no results yet presented, only background, it seems clear that these adjectives of degree refer to the general population. Interestingly, JH annotated this sentence as EXTRANEOUS (0) under Model 2. Sentence 5-3 provides another example; BW and KP annotated it as CONTEXT (1) but JH chose RESULTS COMPARED (4): *Other studies have shown an association between the R219K polymorphism and MI, but no association between haplotype arrangements and MI.* This sentence is problematic for two reasons: there is no citation, and it would seem that these are additional (*other*) studies than *[40]* which is cited in the previous sentence; it contains the first and only mention of *MI* (which I assume stands for 'myocardial infarction', or heart attack) in the entire paper, whereas the previous sentence refers only to *HDL levels*.

Thus given that the current authors make no mention of their results in the context of *MI*, it seems that RESULTS COMPARED (4) is not an appropriate categorization. Another BK~J sentence is 4-4: *Only a small portion of individuals carried these haplotypes, thus the result should be interpreted with caution*, which BW and KP annotated as ANALYSIS (5) and JH as CURRENT RESULTS (3). Although the first clause refers to the patient group in their current study, it is the implication in the second clause that is rhetorically significant; given that ANALYSIS (5) should Trump CURRENT RESULTS (3) if it is "critical to their argumentation", (3) seems an incorrect choice.

Under Model 2 our overall agreement was only slightly worse at 44% than under Model 1, but the ranking was higher, tied for number four (Table 21). BW found this article far more difficult and time-consuming to annotate under Model 2 than under Model 1. This may be related in part to her difficulty in identifying CLAIMS (1) (see above); given that Model 2 is CLAIMS-based, it was hard to identify the authors' lines of argument in the absence of CLAIMS to structure them around. In fact it was her need to find at least one CLAIM (1) that led her to annotate sentence 5-1 as such: *We found an association between LDL levels and ABCA1 genotype, but not with HDL*. In retrospect it seems more a statement of result i.e., GROUNDS (2) (which was JH and KP's annotation) than a "proposition based on analysis and interpretation of results". KP annotated sentence 3-2 as a CLAIM (1): *Ethnic background or other environmental factors may weaken the link with HDL-c levels*. There is no citation here and the authors do not address these 'factors' anywhere in their article, but this sentence is certainly better qualified to be a CLAIM (1) than 5-1. BW and JH both annotated 3-2 as QUALIFIER (4).

The only sentence with total inter-annotator variation under Model 2 was 2-3: *Clee et al also reported lower TG in the carriers of 219K variant and this finding was replicated in our population (p = 0.006)*. Given that the category QUALIFIER (4) includes "compare and contrast with external evidence", BW annotated this sentence as (4); JH identified it as WARRANT/BACKING (3) and KP as GROUNDS (2). Although 2-3 has elements of internal and external evidence, it seems the key rhetorical point is the consistency between these two results. The Trumping guidelines state that QUALIFIER (4) Trumps WARRANT/BACKING (3) where (4) "makes external evidence relevant to their CLAIM"; here it is not clear where this comparison is directed, however, given that there were so few CLAIMS (1). These annotations were in parallel to Model 1 where KP annotated 2-3 as CURRENT RESULTS (3) but BW and JH annotated it as RESULTS COMPARED (4). In a further similarity the four sentences (1-4, 1-5, 3-3, 4-2) with inter-annotator variation between EXTRANEOUS (0) and WARRANT/BACKING (3) under Model 2 all had variation between CONTEXT (1) and either METHOD (2) or RESULTS COMPARED (4) under Model 1 (see above). Sentence 5-3 led to variation under Model 2 as well as Model 1 (see above): BW and KP annotated it as WARRANT/BACKING (3) and JH as QUALIFIER (4), one assumes since he was focused on the "compare and contrast" aspect of QUALIFIER (4). In fact we had the same BK~J annotations (WARRANT/BACKING (3) vs. QUALIFIER (4)) for the block of sentences 5-3 through 5-5. Although only 5-5 has a citation, all three sentences have the current authors reporting on previous studies; as there seems to be no mention in these texts of any current results, and given that category QUALIFIER (4) should make

some reference to the authors' current research, WARRANT/BACKING (3) seems to be the more appropriate category.

3.3.3.11 Article C11 *BMC Infectious Diseases*

The Discussion section of C11 contains 41 sentences with an average length of 23 words (Table 21). This article is unusual in the corpus in that the authors did not perform an experiment, rather they collected and analyzed historical (2004-2006) in-patient medical records, looking at correlations in the data. Thus their Discussion section contains few statements of experimental results, but is focused mainly on work done previously (there are 17 citations) as background or context for their findings, and speculation regarding factors that could account for their results. This latter is typically expressed using the modal hedge *could*, of which there are seven occurrences. This hedge is also used in the Conclusion of the Abstract: *In a critically ill patient with clinical sepsis, GN bacteremia could be associated with higher PCT values than those found in GP bacteremia, regardless of the severity of the disease*. The fact that bacteremia is an 'in vivo' rather than 'in vitro' condition, and *a life-threatening infection* (Background section of the Abstract) suggests that researchers would be especially careful to hedge any non-categorical statements.

Under Model 1 there was relatively good three-way agreement at 68% (28 of 41 sentences), ranking C11 fourth in the corpus; there were no sentences where we all disagreed (Table 22). Given the above regarding the style of C11, it was not surprising that the categories we most frequently all agreed on were CONTEXT (1) with thirteen

sentences and ANALYSIS (5) with nine. In sentence 2-7 we had variation between RESULTS COMPARED (4) (BW and JH) and CONTEXT (1) (KP): *In accordance with our results, some authors have previously shown that in a population with proven sepsis, PCT was significantly higher in patients with bacteremia than in those without [19, 24, 25].* The guidelines state that RESULTS COMPARED (4) would Trump CONTEXT (1), since the opening phrase does provide a significant connection to their current results i.e., it is not simply 'old' information; thus CONTEXT (1) is not an appropriate choice.

Paragraph 4 opens with the statement: *Our findings should, however, be considered with caution.* This is clearly a "limitation of their results" and thus should be annotated as ANALYSIS (5) under Model 1 as BW and JH did. KP's annotation of CURRENT RESULTS (3) is thus incorrect, especially given that ANALYSIS (5) should Trump CURRENT RESULTS (3) if "critical to their argumentation": this statement is a crucial caveat, and sets the stage for the next two sentences. Sentences 4-2 (*First, our results could not be generalized to all patients with sepsis since only those with bacteremia were included.*) and 4-3 (*Secondly, ...is too low to be reliably applied in a clinical setting.*) expand on and give explanations for 4-1. KP annotated these both as METHOD (2), although 4-3 relates their current results to possible applications in the future, and thus METHOD (2) is inappropriate: this category relates only to the methodology of the current study. BW and JH categorized them both as ANALYSIS (5) as they provide implications of their current results for the wider field.

Total inter-annotator agreement was considerably lower under Model 2 at 44% tying it for fourth place with C10, the same ranking as under Model 1 (see Table 21). Eleven of the 41 sentences had variation between EXTRANEOUS (0) and other categories; five of the thirteen BK~J sentences involved JH choosing EXTRANEOUS (0) where BW and KP chose WARRANT/BACKING (3) or PROBLEM IN CONTEXT (5). KP annotated sentence 4-1 as EXTRANEOUS (0) under Model 2; as discussed above, however, this sentence is definitely part of the authors' argument, and should not be categorized as EXTRANEOUS (0). KP annotated 2-7 as QUALIFIER (4) but JH and BW annotated it as WARRANT/BACKING (3); surprisingly we had a reverse of the situation under Model 1: Although KP had CONTEXT (1) (and not RESULTS COMPARED (4)) under Model 1, under Model 2 she had QUALIFIER (4) ("compare and contrast with external evidence"); JH and BW had RESULTS COMPARED (4) under Model 1, but instead of choosing the equivalent under Model 2 of QUALIFIER (4), they both chose WARRANT/BACKING (3), thus not taking the comparison aspect into account. This is a particularly obvious example of 'intra-annotator' variation where one experiences a text differently on different occasions; if this sentence had contained specific data rather than the general *our results* the comparison would have been more striking and inter and/or intra-annotator variation may have been reduced under both Models.

We had three-way variation under Model 2 on sentence 4-9: *However, no patients in our study were given immunosuppressive drugs other than steroids for septic shock. However* connects the discourse to 4-8 which refers to a previous study where *immunosuppressive drugs* might have been used, and which we all annotated as QUALIFIER (4). KP annotated

4-9 as QUALIFIER (4), BW as GROUNDS (2) and JH as EXTRANEOUS (0). Although one could argue for either GROUNDS (2) or QUALIFIER (4), EXTRANEOUS (0) seems inappropriate, especially given that he annotated 4-5 through 4-8 as QUALIFIER (4); if 4-8 is part of the argumentation, then it is hard to see how 4-9 is not. There were only two sentences (3-1 and 3-9) where we had variation between CLAIM (1) and QUALIFIER (4), but this is at least partly because there were not many CLAIMS (1) identified: BW and JH annotated two sentences as CLAIMS (1), and KP four (see Table 31). The small number of CLAIMS (1) may be related to the hedging strategies employed (see above) i.e., the need to provide considerable evidence, internal or external, before making a claim, where serious human disease is being discussed.

3.3.3.12 Article C12 *BMC Molecular Biology*

The Discussion section of C12 has 34 sentences – close to the average of 33 – with an average length of 23 words (Table 21). The first paragraph consists of items the authors wish the reader to consider *in the analysis of [their] results*... [1-1], mainly presenting justifications for their methodology choices. This is a somewhat unusual structure as the opening paragraph typically presents background material to their current study; this defensive position up front regarding their results might be because some of their results are in conflict with what other researchers have found. In sentence 3-1 they state: *The most striking result was the poor performance of some commonly used reference genes* and in the Conclusion section they go even further: ...*questioning the accuracy of previous reports*. This is strong language for the biomedical research community, and they are at pains to buttress their position with 26 citations in the Discussion section. They balance this with more humility in their closing paragraph where the authors focus

on limitations of their experiment, especially concerning their choice of genes and study subjects e.g., ...*we have included a limited array of prospective reference genes.*[5-2], ...*our results only apply directly to*...[5-5]. This argument structure is reflected in the distribution of hedges: except for the use of *assumption* in the first paragraph (in 1-5), all other hedges occur within the final twelve sentences where the authors begin to soften their more challenging positions. In fact C12 has the fewest Table 3 hedges across both corpora, with only six found (see Tables 19 and 33) in 41 sentences.

Under Model 1 three-way inter-annotator agreement in C12 was 65% ranking it fifth in the final corpus (Table 21) and there were no sentences with total disagreement. Six of the twelve sentences with two-way agreement involved variation between CONTEXT (1) and another category. One of these was 4-2: *Best results will be obtained by combining two or three reference genes as emphasized by several authors [30, 32]*, which follows a statement about their current results. KP annotated this as CONTEXT (1), but here the authors cite previous work in order to support their current recommendation to their field; their use of the future tense makes it clear that this aspect of the sentence falls under "suggestions for future work" i.e., ANALYSIS (5). Given the fact that this is a new recommendation, ANALYSIS (5) (JH and BW's annotation) should Trump CONTEXT (1). We had this same JB~K split on sentence 5-4: *Microarray data from cartilage, that now start to be published [5, 10] will provide clues for the identification of the best candidates*. Here, however, the categorization does not seem so straightforward: The current authors are making a statement about the future (*will provide*), but on the basis of

the work of other researchers rather than of their own current results; thus, one could argue for either CONTEXT or ANALYSIS.

Total inter-annotator agreement was extremely low under Model 2 at 21% ranking it at number nine, but only one percentage point better than article C3 at 20% (Table 21). There were five sentences where we all disagreed, four of which involved KP choosing EXTRANEOUS and one where she chose CLAIM (see below). One major source of variation was the first paragraph: As discussed above, BW found all six sentences to be very much a part of the authors' argumentation and saw them as 'qualifying' their results by giving "explanations", and annotated them all as QUALIFIER (4). KP annotated them all as EXTRANEOUS (0) and JH had 1-1 as WARRANT/BACKING (3), 1-4 as QUALIFIER (4) and the balance as EXTRANEOUS (0). One could argue for (3) for all sentences in this paragraph that include citations as providing an "understanding of the problem based on external evidence". We had far less variation under Model 1 as we could agree more easily on the METHOD (2) category. We had two-way variation similar to that under Model 1 on sentences 4-2 and 5-4: KP had WARRANT/BACKING (3) while BW and JH had PROBLEM IN CONTEXT (5). It is worth noting, however, that PROBLEM IN CONTEXT (5) is more specific than ANALYSIS (5) under Model 1: it is clearly future-oriented and related to the results of the current study whereas ANALYSIS can relate to previous work (see Appendix C).

JH annotated no sentence as a CLAIM (1), a problematic situation which will be discussed later in Section 4.2.1.3.3. BW identified three CLAIMS (1), on two of which KP agreed. BW's third CLAIM (1) was in a sentence with three-way variation: *The independence of*

our results from the analysis method gives credence to the conclusions. (2-6) This is certainly a proposition, which may be true or false, but more a stated belief about scientific experimentation than a CLAIM (1) based on analysis of specific results. Thus in retrospect perhaps QUALIFIER (4) (which JH chose) as an "explanation" is the more appropriate categorization. But EXTRANEOUS (KP's annotation) does not seem correct: supporting the credibility of their conclusions is argumentatively significant. In the three other instances of KP's CLAIM annotations, there was variation with QUALIFIER (4) or PROBLEM IN CONTEXT (5).

3.3.4 Hedges

A total of 194 instances of the hedges listed in Table 3, with the minor revisions in Section 2.5.3, were found across the 400 sentences of the 12 articles in the final corpus. Of these the 90 modal verb occurrences account for 46.4% of all hedges. *May* is the most frequently used of all hedges with a total of 35, but note in Table 33 below the particularly high usage in articles C4 (11 times) and C8 (8 times). All nine of the lexical verbs are represented for a total of 67 occurrences, with *suggest* and *indicate* having the most instances. The verb *indicate* however is not evenly distributed across the corpus: eight of the articles have no instances, but article C6 has 11. As in the training corpus, hedging is overwhelmingly accomplished using verbs, with modals and lexical verbs together accounting for 80.9% of all hedge occurrences. Four of the nine possible adjective/adverbs are represented with *possible/possibly* accounting for 16 of the total of 29; together the adjectives and adverbs comprise 15.0% of the hedges. Three of the possible six nouns were identified, accounting for only 4.1% of all hedges. Below are

tables exhibiting hedge distribution by article (with modal verbs highlighted) (Table 33) and by grammatical category (Table 34) in the final corpus:

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Total
Could	4	0	5	0	1	0	1	0	2	0	7	4	**24**
May	0	2	0	11	1	2	1	8	5	3	2	0	**35**
Might	0	0	0	0	2	0	0	0	0	0	2	0	**4**
Should	0	0	3	1	3	0	0	0	0	1	2	0	**10**
Would	0	3	0	2	0	2	4	6	0	0	0	0	**17**
Appear	0	0	0	0	0	2	0	0	2	0	0	0	**4**
Assume	0	0	1	0	1	0	0	0	0	0	0	0	**2**
Believe	1	0	0	0	0	5	0	0	0	0	0	0	**6**
Hope	0	0	0	1	0	0	0	0	0	0	0	0	**1**
Indicate	0	4	1	0	0	11	0	0	4	0	0	0	**20**
Predict	0	0	0	0	0	0	1	0	0	0	0	0	**1**
Seem	1	0	0	0	0	1	0	1	3	0	0	0	**6**
Suggest	6	6	0	3	0	0	2	2	3	3	1	0	**26**
Think/thought	0	0	0	0	0	1	0	0	0	0	0	0	**1**
(Un)likely	2	0	0	0	1	3	0	2	1	0	2	0	**11**
Perhaps	0	0	0	0	0	0	0	0	0	0	1	0	**1**
Possible/y	1	2	0	3	2	1	0	2	3	1	0	1	**16**
Probable/y	1	0	0	0	0	0	0	0	0	0	0	0	**1**
Assumption	0	0	0	0	0	0	0	1	0	0	0	1	**2**
Possibility	0	2	0	0	0	0	0	1	1	0	1	0	**5**
Speculation	0	0	1	0	0	0	0	0	0	0	0	0	**1**
TOTAL	**16**	**19**	**11**	**21**	**11**	**28**	**9**	**23**	**24**	**8**	**18**	**6**	**194**

Table 33: Hedge distribution by article

	Occurrences	Percent
MODAL VERBS	90	46.4%
LEXICAL VERBS	67	34.5%
ADJECTIVES/ADVERBS	29	15.0%
NOUNS	8	4.1%
Total	194	100%

Table 34: Hedge distribution by grammatical category in final corpus

As was the case in the training corpus (Table 19), Table 33 shows a great deal of variation in the use of hedges across the different articles. The average across the corpus

is one hedge per 2.1 sentences (400/194). The most heavily hedged is article C4 with a total of 21 instances in an article of only 24 sentences – averaging almost one hedge per sentence - including 11 uses of *may*. Although C4 also has the greatest number of doubly-hedged sentences (seven), none of these instances involved a hedge collocation. This article presents *unexpected* (3-1) results with implications for their field, thus one would expect the authors to be particularly careful to hedge; however, they also use terms such as *may* and *suggest* extensively when citing previous work, which could mean there is general debate or uncertainty within their particular research specialty (see Section 3.3.3.4 above). At the other extreme is article C12 which has only six hedges in 34 sentences, or one hedge per 5.7 sentences, four of which are *could*. The authors' results seem to challenge the standard approach, going as far as their stating in the Conclusion section that they are *questioning the accuracy of previous reports*; this would suggest a greater rather than minimal use of hedging. As discussed in Section 3.3.3.12 above, however, the authors do make use of non-lexical hedging strategies such as acknowledging a number of limitations of their study.

Hedges were recorded in order of placement within each sentence of an article, but no special indication was made as to whether hedges were collocated. Although it is not uncommon to find hedges collocated together, especially with modal verbs followed by a lexical hedging verb e.g., *this might suggest that, our results may indicate that*, this was not of particular interest in this project; the issue for this study was primarily to identify relationships between individual lexical hedges and categories of argumentation. A sentence is thus counted as 'doubly' hedged whether two hedges in a sentence appear

together in a collocation (as above), or separately within the same sentence e.g., *We believe that...and it seems that....* As presented in Table 35 below, 147 (or 37%) of the 400 sentences in the final corpus contain at least one hedge. 106 sentences have only one hedge, double hedges occur in 35 sentences, and only six sentences contain three hedges ('Triple'). (These (106 + 35*2 + 6*3) total to the 194 instances of hedges, as in Table 33 above.) These two latter categories do not necessarily imply 'stronger' hedging, but are often simply found in longer and/or more complex sentences.

	# of Sentences	# At Least 1 Hedge	Single Hedge	Double Hedge*	Triple Hedge*	Total Hedges
C1	30	13	10	3	0	16
C2	49	16	14	1	1	19
C3	25	9	7	2	0	11
C4	24	14	7	7	0	21
C5	33	10	9	1	0	11
C6	45	19	12	5	2	28
C7	21	7	5	2	0	9
C8	35	16	11	3	2	23
C9	36	17	10	7	0	24
C10	27	6	5	0	1	8
C11	41	14	10	4	0	18
C12	34	6	6	0	0	6
Total	**400**	**147**	**106**	**35**	**6**	**194**

Table 35: Distribution of hedges within sentences by article
* Hedges are within a sentence but not necessarily collocated

Under Model 1 there is a higher probability of total annotator agreement on the categorization of sentences that contain a hedge than of all sentences in the corpus: We all agreed on argument category for the sentences containing 143 (73.7%) of the 194 instances of hedging, whereas we had total agreement on the categorization of only 60.5% of all 400 sentences in the corpus under Model 1 (see Table 22). Note that I refer to "instances" of hedging rather than hedged sentences as a single sentence may contain

more than one lexical hedge: there are 194 hedge occurrences, but only a total of 147 hedged sentences.

There was two-way agreement on argument category for sentences containing 47 (24.2%) of the 194 instances of hedging, compared to 35.8% of all 400 sentences. We had total disagreement on argument category for sentences containing only four (2.1%) of the 194 instances of hedging. The data for the three levels of annotator agreement on sentences containing hedges under Model 1 are displayed below in Table 36:

	ALL AGREE	2-WAY SPLIT	ALL DISAGREE	TOTAL
Could	18	4	2	**24**
May	28	7	0	**35**
Might	3	1	0	**4**
Should	7	3	0	**10**
Would	6	10	1	**17**
Appear	2	2	0	**4**
Assume	2	0	0	**2**
Believe	6	0	0	**6**
Hope	1	0	0	**1**
Indicate	9	10	1	**20**
Predict	1	0	0	**1**
Seem	5	1	0	**6**
Suggest	24	2	0	**26**
Think/thought	1	0	0	**1**
(Un)likely	9	2	0	**11**
Perhaps	1	0	0	**1**
Possible/y	14	2	0	**16**
Probable/y	0	1	0	**1**
Assumption	2	0	0	**2**
Possibility	4	1	0	**5**
Speculation	0	1	0	**1**
TOTAL	**143**	**47**	**4**	**194**
	73.7%	**24.2%**	**2.1%**	**100%**

Table 36: Inter-annotator agreement groupings for hedged sentences – Model 1

There was considerably less three-way inter-annotator agreement on the categorization of sentences that contain a hedge under Model 2 than Model 1, but this reflected the overall difference in inter-annotator agreement between the two Models for all sentences in the corpus (see Tables 22 and 23). We all agreed on argument category for sentences containing 68 (35.1%) of the 194 hedges under Model 2, whereas we had total agreement on 39.3% of the 400 sentences in the corpus. There was two-way agreement on argument category for sentences containing 110 (56.7%) of the 194 instances of hedging under Model 2, slightly more than the 52.5% two-way agreement on all sentences across the corpus. We all disagreed on argument category for sentences containing 16 (8.2%) of the 194 instances of hedging, similar to the 8.3% of all 400 sentences with total disagreement under Model 2 (Table 23). The data for the three levels of annotator agreement on sentences containing hedges under Model 2 are displayed below in Table 37:

	ALL AGREE	2-WAY SPLIT	ALL DISAGREE	TOTAL
Could	5	15	4	**24**
May	10	23	2	**35**
Might	2	2	0	**4**
Should	2	7	1	**10**
Would	5	10	2	**17**
Appear	3	1	0	**4**
Assume	1	1	0	**2**
Believe	1	5	0	**6**
Hope	1	0	0	**1**
Indicate	10	9	1	**20**
Predict	0	1	0	**1**
Seem	1	3	2	**6**
Suggest	11	14	1	**26**
Think/thought	0	1	0	**1**
(Un)likely	5	5	1	**11**
Perhaps	1	0	0	**1**
Possible/y	6	9	1	**16**
Probable/y	0	0	1	**1**
Assumption	1	1	0	**2**
Possibility	3	2	0	**5**
Speculation	0	1	0	**1**
TOTAL	**68**	**110**	**16**	**194**
	35.1%	**56.7%**	**8.2%**	**100%**

Table 37: Inter-annotator agreement groupings for hedged sentences – Model 2

Since not all annotators agreed on the choice of argument category for the sentences containing hedges (26.3% of the hedged sentences under Model 1 and 64.9% under Model 2 have some degree of inter-annotator variation), the argument categorization statistics in Tables 38 and 39 below are based on the choices of individual annotators. Thus, the total number of tokens for category annotations in hedged sentences is 582: 194 times three annotators. Under Model 1 sentences containing hedges are overwhelmingly categorized as ANALYSIS (5): 432 (74.2%) of the 582 possible hedged sentence categorizations. The next most frequent category is CONTEXT (1) with 86 (14.8%) of the

sentences containing hedges. The remaining categories are not well represented: only 16 for METHOD (2), 25 for CURRENT RESULTS (3) and 23 for RESULTS COMPARED (4). The distribution of hedged sentence annotations by Model 1 category is displayed below in Table 38:

HEDGE	CONTEXT (1)	METHOD (2)	CURRENT RESULTS (3)	RESULTS COMPARED (4)	ANALYSIS (5)	Total
Could	7	1	0	2	62	**72**
May	12	1	1	2	89	**105**
Might	0	0	0	1	11	**12**
Should	11	1	2	0	16	**30**
Would	12	6	4	1	28	**51**
Appear	2	1	2	1	6	**12**
Assume	3	0	0	0	3	**6**
Believe	0	0	0	3	15	**18**
Hope	0	0	0	0	3	**3**
Indicate	13	0	14	10	23	**60**
Predict	0	0	0	0	3	**3**
Seem	0	2	0	1	15	**18**
Suggest	13	0	0	1	64	**78**
Think/thought	0	0	0	0	3	**3**
(Un)likely	4	0	1	0	28	**33**
Perhaps	0	0	0	0	3	**3**
Possible/y	3	1	0	0	44	**48**
Probable/y	1	0	0	0	2	**3**
Assumption	0	3	0	0	3	**6**
Possibility	3	0	1	0	11	**15**
Speculation	2	0	0	1	0	**3**
TOTAL	**86**	**16**	**25**	**23**	**432**	**582**
PERCENT	**14.8%**	**2.7%**	**4.3%**	**4.0%**	**74.2%**	**100%**

Table 38: Distribution of hedged sentence annotations by category – Model 1

Under Model 2 the two most common categories for sentences containing hedges are QUALIFIER (4) with 207 (35.6%) of the total 582 tokens, and CLAIM (1) with 183 (31.4%).

In decreasing frequency, the remaining categories are EXTRANEOUS (0) with 69 (11.9%), WARRANT/BACKING (3) with 54 (9.3%), PROBLEM IN CONTEXT (5) with 40 (6.9%) and GROUNDS (2) with 29 (4.9%) of all 582 tokens. The distribution of hedged sentence annotations by Model 2 category is presented below in Table 39:

HEDGE	EXTRANEOUS (0)	CLAIM (1)	GROUNDS (2)	WARRANT/ BACKING (3)	QUALIFIER (4)	PROBLEM IN CONTEXT (5)	Total
Could	10	16	0	5	26	15	**72**
May	11	33	2	7	43	9	**105**
Might	0	3	0	0	9	0	**12**
Should	11	2	5	0	9	3	**30**
Would	14	12	1	3	19	2	**51**
Appear	0	2	4	3	3	0	**12**
Assume	1	3	0	2	0	0	**6**
Believe	0	10	0	1	7	0	**18**
Hope	0	0	0	0	0	3	**3**
Indicate	0	19	16	16	9	0	**60**
Predict	0	1	0	0	0	2	**3**
Seem	1	9	1	2	5	0	**18**
Suggest	6	47	0	9	16	0	**78**
Think/thought	0	1	0	0	2	0	**3**
(Un)likely	5	7	0	4	16	1	**33**
Perhaps	0	0	0	0	3	0	**3**
Possible/y	5	13	0	0	25	5	**48**
Probable/y	1	1	0	0	1	0	**3**
Assumption	2	3	0	0	1	0	**6**
Possibility	0	1	0	2	12	0	**15**
Speculation	2	0	0	0	1	0	**3**
TOTAL	**69**	**183**	**29**	**54**	**207**	**40**	**582**
PERCENT	**11.9%**	**31.4%**	**4.9%**	**9.3%**	**35.6%**	**6.9%**	**100%**

Table 39: Distribution of hedged sentence annotations by category – Model 2

In Section 1.4 I hypothesized that hedges would be likely to occur in sentences annotated as QUALIFIER (4) under Model 2, given that this category applies where authors present

possible or probable explanations for their findings. The final corpus results above show that (4) is the most frequent category at 35.6%, but that the category CLAIM (1) accounts for almost as many hedged sentence annotations – 31.4%. The inter-annotator variation between these two categories was both frequent and problematic suggesting that it was sometimes difficult for annotators to differentiate between them; this variation will be discussed in detail in Section 4.2.1.2. This split between CLAIMS (1) and QUALIFIER (4) is in marked contrast to Model 1 where our hypothesis in 1.4 was strongly confirmed: sentences containing hedges were annotated as ANALYSIS (5) almost 75% of the time (Table 38). Analysis and implications under both Models of the above results regarding hedges can be found in Section 4.3.

3.3.5 Argument Type

The list of Argument Types used in the final corpus was presented in Table 20 (Section 2.5.4). As mentioned there, Type (4) – Leads to/opens up new research direction, refines an existing research question, or contributes to addressing a significant research issue – is of a more general nature than the first three. Given this, I hypothesized that it might be more commonly selected than Types (1) Advanced/improved methodology or experimental design, (2) New creation/concept, or (3) New way of looking at/interpreting/evaluating existing data/previous results. As displayed in Table 40 below, however, this was true only for BW. Again, the fact that BW chose (4) for half of the articles in the final corpus, whereas JH used it only once, and KP not at all, may be due to her lack of expertise in the scientific content. JH and KP may have been better equipped to recognize the first three Types i.e., the more specific core aspects of novelty which the

researchers were presenting. Overall the most commonly selected Argument Types were (2) and (3):

ARGUMENT TYPE	BW	JH	KP	TOTAL INSTANCES
1	2	2	2	6
2	3	4	5	12
3	1	5	5	11
4	6	1	0	7
TOTAL	12	12	12	36

Table 40: Distribution of Argument Types by annotator in final corpus

As shown in Table 41 below our inter-annotator variation on Argument Type was generally of the two-way split variety: we had two-way agreement on nine of the twelve articles (75%) in the final corpus (2 were JK~B, 2 were BK~J, and 5 were JB~K). This skew toward JH and BW agreeing where KP differs was not reflected in our overall annotations of argument category under Model 1 (where KP and BW were more likely to agree) or Model 2 (where all two-way agreement groupings were equally distributed) (see Tables 22 and 23). The subjectivity and difficulty of this task are reflected in the fact that the three annotators did not agree on Argument Type for a single article, but all disagreed on three (25%): C2, C9 and C12.

Annotator	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
BW	3	4	2	2	1	4	4	1	4	4	2	4
JH	3	1	2	3	1	4	3	2	3	3	2	2
KP	2	3	3	2	2	2	3	1	2	3	3	1

Table 41: Argument Type by article and annotator in final corpus

BW chose Type (4) for all three articles where there was total disagreement; as discussed earlier, BW found the scientific content of C2 and C9 to be the most inaccessible of the

corpus. All three annotators found that article C12 was not clearly written, and thus difficult to assess under either Model of argument. The notion of Argument Type was that it is inherent to the argumentation of, and general enough to capture the core argumentative purpose in, any *BMC* article (Discussion section). As has been shown in Section 3.3.3 above, however, the articles in our final corpus are complex, and do not seem to fit neatly into (only) one of the four Types in Table 20; nevertheless, the amount of inter-annotator variation here was still surprising. The issue of Argument Type is discussed further in Section 4.4.

CHAPTER 4 DISCUSSION and ANALYSIS

4.0 Introduction

In this Chapter I begin by comparing inter-annotator agreement in the training corpus (Chapter 2) and the final corpus (Chapter 3) in Section 4.1. Average three-way agreement did not improve between the two corpora under Model 1 but was higher in the final corpus under Model 2; average three-way disagreement, however, was reduced under both Models. I also note that there was a much wider range between the lowest and highest three-way agreement values among the articles in the final corpus, under both Models of argument. In Section 4.2 I discuss inter-annotator variation and its sources: I begin by looking at each Model in turn, providing data on the most frequent inter-category disagreements (Section 4.2.1). I then look at pair-wise inter-annotator variation, including crosstabulations by Model and category for each annotator pairing in the final corpus. I also discuss errors made by annotators, including apparent misunderstandings of Model categories or argument structure, as well as 'performance errors' (Section 4.2.1.3). In Section 4.2.1.4 I discuss the corpus data as a source of variation; this covers the problems of complex sentences (common in all texts), technical biomedical content (especially for non-experts), and also the range in writing skills and styles across articles (inter-article variation). Next I address the issue of hedges, comparing the frequency distributions across the two corpora, and relating hedge occurrences to argument category under both Models (Section 4.3). Finally in Section 4.4 I briefly discuss the results on Argument Type: despite a complete revision of the list of Types following discussions during training, there was no three-way inter-annotator agreement on Argument Type in the twelve articles in the final corpus.

4.1 Inter-annotator Agreement

It is not possible to draw detailed comparisons between the results of the training annotations in Chapter 2 and those of the final corpus in Chapter 3 given the following factors: The process with articles T1 and T2 provided an introduction to the concepts and data of the project for JH and KP, and thus results were very preliminary. Annotation units smaller than the sentence were allowed for T1 and T2 only, thus these results are not congruent with the balance of the project. In addition, Model 2 was new to BW and she was still being trained herself to become familiar with its theory and categorizations. The same Models of argument were applied across all five training articles, but the sentence was enforced as the unit of annotation for T3 through T5. Before the annotation of the final corpus both Models 1 and 2 were revised, and Trumping guidelines were added to each (Appendices C and D). It is not possible therefore to compare argument categorizations between the training and final corpora.

As presented in Tables 22 and 23 total inter-annotator agreement occurred in 60.5% of the final corpus sentences under Model 1 and 39.3% under Model 2. This implies that there was some type of inter-annotator variation in 39.5% of the sentences under Model 1 and 60.7% under Model 2. Thus, we have essentially mirror images of the agreement/variation distribution between the two Models. Although the amount of systematic variation is too high to allow us to use an agreement coefficient such as the Kappa Statistic (Carletta 1996), it is intuitively clear that there are problems, especially with Model 2, when we see that in article C4 all annotators agreed on only two of the 24

sentences. Although Model 1 produced higher overall agreement results, and its lowest three-way agreement percentage was 36% (article C5), only 3.3% below the *average* for Model 2, the fact remains that there was virtually no improvement in the average three-way inter-annotator agreement between the first training phase and the final corpus (Tables 18 and 21). The facts that with both Models the average three-way inter-annotator agreement is relatively low, and that there is such an extreme range of overall agreement among articles, implies that there are problems with the Models and/or the annotators in terms of consistency. In order to improve agreement, therefore, it is crucial to attempt to analyze the variation found across both corpora, and to identify the sources of the variation. The issue of variation will be discussed in detail in Section 4.2 below.

4.1.1 Model 1

Despite the above limitations however, one of the key goals of this project was to see if the relatively lengthy training phase would lead to improved inter-annotator agreement in the final corpus. In fact, based on the averages for Model 1 this was not the case; it is striking that for all three corpus groups identified below in Table 42 the average three-way inter-annotator agreement was virtually identical at around 60%. The percent of sentences on which we all disagreed was reduced from the high of 8.1% for the first two training articles, averaging 3.8% by the final corpus, a sign of some increase in our shared understanding of the Model 1 categories following phase I. Interestingly the average percentage of sentences where JH and BW agreed was cut almost in half in the final corpus, whereas those where BW and KP agreed increased dramatically from an average of 5.6% during training to 17.3% in the final annotations. The percent of JK~B

sentences for T3-T5 (shaded below) is somewhat inflated, being skewed by the results for article T3 (see Table 13), where BW found the content extremely difficult, as discussed in Section 2.3.2.2.1. Agreement groups for Model 1 are found in Table 42 below[4].

	T1-T2	T3-T5	C1-C12	Average
# Sentences	62	85	400	
All agree	61.3%	58.8%	60.5%	60.2%
All disagree	8.1%	2.4%	3.8%	4.8%
JK~B	4.8%	15.3%	8.0%	9.4%
JB~K	19.4%	18.8%	10.5%	16.2%
BK~J	6.4%	4.7%	17.3%	9.5%

Table 42: Inter-annotator agreement groups in training and final corpora – Model 1

The averages for overall inter-annotator agreement above mask a significant difference between the training and final corpora: the range among articles T1-T5 was only 21% (between 50% (T3) and 71% (T2), see Table 18) whereas for C1-C12 it was more than twice that at 45% (between 36% (C5) and 81% (C7), see Table 21). The smaller sample size (147 sentences vs. 400 in the final corpus) may account for part of the former: given enough data, the probability of encountering 'easier' and 'more difficult' articles would increase, and our range in agreement would ultimately increase, as it did in C1-C12. Note that the articles used in the training phase were not selected based on their apparently lower level of difficulty from an annotation perspective; all 17 articles were selected equally randomly, with the assumption that they are a representative sample from the *BMC*-series of journals. During the training period, however, discussions and questions between annotators were allowed, which likely led to a higher level of agreement in the

[4] Note that here and elsewhere there may be minor variations in percent statistics (e.g., for overall inter-annotator agreement) between different tables; these are a consequence of differing calculations, depending on the variable(s) being considered, which lead to different 'rounded' results. Thus in Tables 42 and 43 below 'rounding errors' lead to the averages totalling 100.1%.

training results than if annotators had simply been given the printed instructions with no opportunity for consultation.

In addition, revisions were made to the unit of annotation and to Model 1 before the annotation of the final corpus, so I am unable to make direct comparisons between the categories on which all annotators agreed during the training (Tables 9 and 15) and final (Table 24) corpora. It is worth noting here that the new category CONTEXT (1), although added to simplify Model 1 and improve agreement, was a major source of inter-annotator variation in the final corpus (see Table 46 below for detailed data). Although it was a category with almost 25% of the total agreement sentences (Table 24), there was an overall bias against this category and toward RESULTS COMPARED (4) on the part of JH (see Table 27). This asymmetry with BW and KP was a major factor in the dramatic increase in the percent of BK~J sentences in the final corpus noted above. Concomitantly, there was a decrease in the JB~K sentences in the final corpus to 10.5%, although no single factor is evident here.

The most noticeable improvement in the final corpus is found at the high end of the three-way agreement range: two articles (C8 and C1) had 80% total agreement, and C7 had 81% (Table 21), whereas during training the highest level was 71% in T2 (Table 18). On the other hand, although the lowest three-way agreement in the training data was 50% in T3 (Table 18), in the final corpus we had three articles with lower agreement: C5 at 36%, C3 at 44% and C6 at 47% (Table 21). So despite the wider range, the average three-way inter-annotator agreement, as shown in Table 42, remained essentially unchanged. There

is no clear answer as to why the average did not improve after training. As mentioned earlier, JH and KP were relatively familiar from the outset with the information-based concepts behind Model 1, so the training process may not have been as crucial as for Model 2. For JH and KP the training process involved the attention warranted by employees working on a new task as well as the detailed written feedback on their processes which were required; on the other hand, the final annotations were done independently and in isolation, where the only feedback was their annotated data. This is not to suggest that either JH or KP were careless in their final annotations, but only that without supervision they were not able to seek clarification if they had questions. It also seems to be the case that the revisions to Model 1, including the addition of Trumping guidelines, resolved some problems but created new sources of variation; Trumping will be discussed further in Section 5.2.

Perhaps the most striking observation of all is the wide range of three-way agreement among the 12 articles of the final corpus (Table 1) as discussed above. I was surprised by this finding, and asked myself the following question: How to account for this when the same three annotators had the same training and used the same Model 1 argument categories and Trumping instructions? It seems clear that the content of the articles – style of writing and argumentation, as well as issues related to the biomedical field and/or the particular experiment – is an important factor in this range of inter-annotator agreement (see Section 4.2.1.4 below). The detailed discussions in Sections 2.3.2.2.1-3 and 3.3.3.1-12 attempt to address the particular characteristics of each article, thus

providing a survey of the different types of data which can be found in the *BMC*-series of journals and offering possible explanations for specific annotation results.

4.1.2 Model 2

Unlike under Model 1, we do see some improvement in the average three-way inter-annotator agreement between the training and final corpora under Model 2: from 32.2% to 39.3%. There is also a gradual and notable decrease in the average three-way disagreement from 21.0% to 8.3% (see Table 43 below). The average percentage of two-way agreement sentences in the final corpus (52.6%) was virtually the same as that in the training corpus (51.6%), although the distribution among the three sub-categories varied somewhat (see below). As under Model 1 above, the percentage of JK~B sentences (shaded below) for T3-T5 was unusually high based on the skewed distribution of article T3. Inter-annotator agreement groupings for the training and final corpora under Model 2 are displayed below in Table 43:

	T1-T2	T3-T5	C1-C12	Average
# Sentences	62	85	400	
All agree	32.3%	31.7%	39.3%	34.4%
All disagree	21.0%	11.8%	8.3%	13.7%
JK~B	17.7%	27.1%	17.8%	20.9%
JB~K	12.9%	12.9%	17.0%	14.3%
BK~J	16.1%	16.5%	17.8%	16.8%

Table 43: Inter-annotator agreement groups in training and final corpora – Model 2

In parallel with Model 1 the averages of the three-way agreement percentages above give no indication of the range across articles of the corpora. It is significant to note that this agreement ranged from 22% (T5) to 42% (T2) in the training corpus and from 8% (C4) to 69% (C6) in the final corpus (Tables 18 and 21). The former spread of 20% is virtually identical to that for training under Model 1, but the latter range at 61% is considerably higher than the 45% under Model 1. If C4 is removed as an outlier, however, the final corpus range for Model 2 is reduced to 49% (from 20 (C3) to 69%). Possible explanations for the difference in spread between the training and final corpora are similar to those for Model 1 above i.e., smaller vs. larger corpus, and annotating with consultation vs. working in isolation. And as discussed above for Model 1, it is important to investigate the stylistic and argumentative qualities of the individual articles in order to try to account for the wide range of overall inter-annotator agreement within and across the corpora.

As with Model 1 I am unable to draw direct comparisons between categories on which we all agreed in training (Tables 10 and 16) and the final corpus (Table 25) as these categories underwent revisions. Nevertheless there was a notable shift between the training and final corpora in terms of CLAIMS: in T1-T5 BW identified 12 CLAIMS, JH 29 and KP 23; in C1-C12 the distribution was 45 for BW, 49 for JH and 91 for KP (see Table 44 below). JH's bias away from CLAIMS in the final corpus along with KP's propensity to identify them may account for at least some of the increase from 12.9 to 17.0% in JB~K sentences in Table 43. The training process also seemed to clarify to some degree our understanding of what constitutes a CLAIM as the three-way agreement

on this category improved in the final corpus (see Table 44 below). Note that this total of 24 includes five sentences in article C6, the one article with higher three-way agreement than under Model 1.

	# Sentences	**BW**	**JH**	**KP**	**All agree on CLAIM**
T1-T2	62*	6	16	12	0
T3-T5	85	6	13	11	1
C1-C12	400	45	49	91	24
Total	547	57	78	114	25

Table 44: Number of CLAIMS identified by annotators across corpora and number with three-way agreement

*sub-sentential annotation units allowed

Between the training and final corpora there was an increase of approximately 7% in the average three-way inter-annotator agreement (see Table 43). Given that Model 2 was unfamiliar to all three annotators at the beginning of this project, and that BW was still training herself while annotating articles T1-T5, it is not surprising that there would be some improvement in overall agreement following the training process. Certainly consultations with Graves during training helped to increase our understanding of Model 2, a fact which I am sure reduced our level of inter-annotator disagreement. In addition, revisions were made to the Model, including the addition of Trumping guidelines (Appendix D). It is not possible to know which of these factors played the main role in the improvement in overall agreement.

Despite this improvement, however, the average three-way inter-annotator agreement level remained low at 39.3% in the final corpus, and more than 20% lower than the

average for Model 1. The differences in overall agreement between Models 1 and 2 are shown for each article in the final corpus in Table 45 below. Reflecting the wide range of variation in three-way agreement under both Models, Model 1's performance ranges from being 57% higher than that for Model 2 in article C1 to 22% lower in article C6 (shaded; the sole article where Model 1 had less overall agreement than Model 2).

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
M1	80%	53%	44%	58%	36%	47%	81%	80%	69%	52%	68%	65%
M2	23%	51%	20%	8%	33%	69%	62%	34%	39%	44%	44%	21%
M1 vs. M2	+57%	+2%	+24%	+50%	+3%	-22%	+19%	+46%	+30%	+8%	+24%	+44%

Table 45: Comparison of total inter-annotator agreement between Models 1 & 2 in final corpus

Although Sections 3.3.3.1-12 present speculation regarding reasons for the above variation between Models in each particular article, there are two factors which apply generally across the corpus: a) all three annotators were less familiar with the concepts behind Model 2 and thus found it more difficult to be certain in mapping from the category descriptions to corpus data, and b) Model 2 has six categories, one more than Model 1, and therefore more possibility for variation. In addition, Model 2 had considerable inter-annotator variation between categories (1) CLAIM and (4) QUALIFIER, a conflict which had no real analogue under Model 1. This and other sources of variation will be discussed in detail in Section 4.2.1.2 below.

4.2 Inter-annotator Variation

As discussed above, given that agreement and variation are essentially in an inverse relation to each other, I cannot talk about one without addressing the other. Even the term "two-way agreement" necessarily implies "two-way variation": where two annotators agree on a categorization and the third disagrees, there is variation between two different argument categories. Given that there was no 'expert' annotator used in this project (i.e., the standard against which other annotators' results would be compared), and that there are numerous instances where more than one annotation for the same sentence could be deemed acceptable, I view the majority of inter-annotation variation as 'legitimate'. By this I mean that it is worthy of analysis and study with the goals of evaluating and improving the Models of argument, as well as gaining insight into who might be the 'ideal' annotator for these biomedical texts. The exception to this would be outright 'errors' in the annotations, some of which have been identified in Sections 2.3.2.2.1-3 and 3.3.3.1-12; this variation is a reflection of problems with annotators, rather than the Models of argument. Annotator errors will be discussed below in Section 4.2.1.3.3.

I have referred earlier to 'intra-annotator variation', where a single annotator may provide a different annotation for the same unit of text over a period of time. This is a phenomenon BW experienced during previous annotations, where it was generally the result of a better understanding of the article's content on rereading, rather than of the model of argument being applied (White 2005a). During the training process for the current project, consultation and revisions to annotations were permitted. Since all annotators worked independently on the final corpus, however, there was no control over

e.g., how much time was spent on reading or annotating, if multiple articles were annotated in quick succession, whether annotations were revised over time before submission, etc. No changes were made to the final corpus annotations once the data were collected from the three annotators, despite BW's considering in retrospect during data compilation and analysis that she would have made some revisions. Thus we do not consider intra-annotator variation in the statistics describing our results.

Again I return to the position that it is impossible to isolate single causes of inter-annotator variation in these results; there are numerous sources, and interactions among them are the rule rather than the exception. This is especially true given that this was a small-scale pilot project where the goals were primarily to assess the ease of application and utility of the two Models of argument, and secondarily to investigate the appropriateness of the three annotators for this task; in other words, I expected variation and wanted to explore it. It is by identifying not only the dimensions along which the inter-annotator variation occurred, but also where these factors appeared to intersect, that we hope to develop better Models of argument (e.g., with categories more clearly differentiated, with more definitive directions re Trumping, etc.) and improve the level of inter-annotator agreement. I have touched on and exemplified many of the problems encountered through this project in Sections 2.3.2.2.1-3 and 3.3.3.1-12 by discussing individual articles in turn. In the next Sections I summarize the key sources of variation and their impact on the annotation results: the Models of rhetoric, the annotators and the corpus data.

4.2.1 Sources of variation

4.2.1.1 Model 1

As discussed above, there was some degree of inter-annotator variation in 39.5% of the sentences under Model 1 in the final corpus. The fact that there was three-way disagreement on only 3.8% of the sentences suggests that there was some degree of shared understanding by all annotators of the Model 1 categories. In fact in some of these (e.g., sentences 5-2 and 6-3 in C2 (Section 3.3.3.2), among others) the three different category choices all seemed legitimate i.e., the result of subjective differences rather than errors; although the number of such sentences is relatively small, they could be indicative of a need to make the category definitions more precise.

The new category (1) CONTEXT was by far the major source of inter-annotator variation in the final corpus. As presented in Section 2.5.1, the original Model 1 category (2) Issues under Dispute (Table 1) was the cause of considerable confusion and variation during training annotations and was therefore eliminated from the original Model 1. The CONTEXT category was developed to separate out 'old' from 'new' information in an article, to include content that provides 'context' to the current experiment (Appendix C). The original category (1) Previous Work (Table 1) was rolled into the new CONTEXT (1) category.

Ample examples of this inter-annotator variation between CONTEXT (1) and other Model 1 categories have been given in Sections 3.3.3.1-12 above; here I provide the distribution of this variation among specific categories and across articles. The most striking

observation is that the variation involving CONTEXT (1) occurs in almost two thirds of all sentences with some level of variation: 100 out of 158 sentences (see Table 46 below)[5]. The problem at the core of this variation seems to be that researchers are always working within and building on the 'context' of previous work – their own and others'; thus statements about an experiment are likely to be enmeshed in statements referencing other work, including previous methodologies and results. This may take the form of a complex sentence, leading to difficulty when annotators must choose a single category per sentence. Although the addition of the applicability of category (5) ANALYSIS to previous as well as current results may have reduced some variation (44.6% of the three-way agreement sentences were category (5)), it introduced the increased possibility of variation between (1) and (5). As shown in Table 46 below, 39 such sentences occurred in the final corpus:

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Total
# Sentences	30	49	25	24	33	45	21	35	36	27	41	34	400
# with some Variation	6	23	14	10	21	24	4	7	11	13	13	12	158
# with Variation involving CONTEXT (1)	4	17	6	7	19	9	3	5	7	9	8	6	100
Some (1) vs. (2) METHOD	0	4	0	0	10	1	0	1	3	2	3	2	26
Some (1) vs. (3) CURRENT RESULTS	0	2	1	0	1	4	1	0	0	0	0	1	10
Some (1) vs. (4) RESULTS COMPARED	2	6	3	0	3	5	1	3	4	5	3	1	39
Some (1) vs. (5) ANALYSIS	4	11	2	7	14	0	1	2	2	2	2	2	53
Total 4 rows above	6	23	7	7	20	10	3	6	9	9	8	6	114

Table 46: Inter-annotator variation between CONTEXT (1) and other categories – Model 1

[5] There is some overlap between the variation groupings, thus they add up to 114 although the number of sentences with some variation involving CONTEXT (1) totals only 100.

Variation between (1) and RESULTS COMPARED (4) (in 39 sentences) is not surprising given that category (4) by definition is referencing previous results; in contrast we see only 10 sentences with variation between (1) and CURRENT RESULTS (3). The total number of sentences with variation between (1) and METHOD (2) appears surprisingly high at 26, but this is inflated by the 10 instances found in C5, the only paper listed as a "Methodology article". As discussed in Section 3.3.3.5 this article had the lowest inter-annotator agreement under Model 1, and 19 of the 21 sentences with some variation involved the CONTEXT (1) category. Numerous sentences describing background facts and citing previous work are at the same time discussing, explaining and defending their current methodology.

The implication of these results seems to be that trying to separate out 'old' from 'new' information in biomedical research texts, at least where the sentence is the unit of annotation, is not productive; trying to make this distinction in fact generates, rather than reduces, inter-annotator variation. Since the Trumping guidelines use the same basis of new vs. historical information (e.g., categories (2)-(5) Trump (1), Appendix C), and as it applies only where an annotator feels conflicted, it does not seem to have been successful in reducing the variation evident in Table 46. Given that 'old' and 'new' are inextricably linked in biomedical research, the CONTEXT (1) category needs to be revised or eliminated from Model 1.

As discussed in Section 2.5.1, after the training phase it was decided to revise category (5) ANALYSIS so that it could refer to either current or previous results (see Appendix C).

It is believed that this clarification made this category easier to agree on overall: it accounted for 44.6% of the sentences with three-way agreement in the final corpus (Table 24) although only 36.0% of all category tokens (Table 26). Nevertheless, since ANALYSIS is the most frequent overall category choice it is not surprising that it is also involved in considerable inter-annotator variation. This revision to Model 1 thus appears to have reduced some inter-annotator variation in the final corpus, but as shown above, the introduction of the CONTEXT (1) category added significantly to the overall variation involving ANALYSIS. (Although as stated earlier I cannot make one-to-one comparisons between the training and final results, I refer the reader to Tables 9 and 15 for an overview of the distributions of categories with total agreement under Model 1 in the training corpora.)

In Table 47 below I present the number of sentences involving some variation between ANALYSIS (5) and other categories. I repeat the data from Table 46 for CONTEXT (1), totalling 39 sentences. Variation with the categories METHOD (2), CURRENT RESULTS (3) AND RESULTS COMPARED (4) total 41, but only 39 distinct sentences are involved. Note that these may also include some variation with CONTEXT, so there will be overlap between these 39 sentences and the 39 for CONTEXT (1).

I also include in Table 47 the number of sentences with some variation between the categories (3) CURRENT RESULTS and (4) RESULTS COMPARED; these total 16, although only six of the twelve articles contained any of these sentences. It is not known whether the Trumping guidelines prevented these inter-annotator variation results from being

higher as there was no feedback from annotators on where they did or did not make use of them in the final corpus. (Any possible inter-category variations other than those in Tables 46 and 47 were non-existent or of a trivial number.)

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Total
# Sentences	30	49	25	24	33	45	21	35	36	27	41	34	400
# with some Variation	6	23	14	10	21	24	4	7	11	13	13	12	158
Some ANALYSIS (5) vs. CONTEXT (1)	4	11	2	7	4	0	1	2	2	2	2	2	39
Some ANALYSIS (5) vs. METHOD (2)	0	4	2	0	2	0	0	0	0	0	2	1	11
Some ANALYSIS (5) vs. CURRENT RESULTS (3)	1	1	1	0	1	1	0	1	3	1	2	1	13
Some ANALYSIS (5) vs. RESULTS COMPARED (4)	1	2	1	3	1	2	1	2	4	0	1	0	18
Some CURRENT RESULTS (3) vs. RESULTS COMPARED (4)	0	4	4	0	0	3	0	0	1	3	0	1	16

Table 47: Inter-annotator variation between categories – Model 1

4.2.1.2 Model 2

Before examining variation between categories under Model 2 I reiterate two key differences with Model 1: Model 2 has approximately 20% more inter-annotator variation than Model 1 and has one more argument category than the five in Model 1. The categories primarily involved in variation under Model 2 were EXTRANEOUS (0), especially vs. WARRANT/BACKING (3); and QUALIFIER (4), especially vs. CLAIM (1). Both EXTRANEOUS (0) and QUALIFIER (4) were problematic for reasons that will be discussed below, but they are also the most frequently occurring categories in the corpus at 20.8% and 21.3% respectively (Table 29); thus, aside from the particular difficulties with the specifications for categories (0) and (4), their frequency makes them more likely to

appear in sentences with inter-annotator variation. Variation between the following categories will not be discussed here as they represent very few sentences in the corpus, and some are the result of annotator error: GROUNDS (2) vs. WARRANT/BACKING (3); GROUNDS (2) or WARRANT/BACKING (3) vs. PROBLEM IN CONTEXT (5).

As discussed in Section 2.5.2 the category EXTRANEOUS (0) was added to Model 2 to incorporate statements that are not part of the authors' line of argumentation: this would include material external to Toulmin's argument structure, as presented in Section 1.3.2.1, which essentially uses evidence to support a CLAIM. Although there was some overlap with the specifications for the CONTEXT category in Model 1 e.g., background material, at a theoretical level it was different: CONTEXT was designed to indicate statements of 'old' information whereas EXTRANEOUS was based on argument structure rather than timing. Nevertheless, in some parallel with Model 1, especially involving statements about research done previously, the introduction of this new category was the major source of inter-annotator variation under Model 2.

Some of the difficulties with deciding which statements are EXTRANEOUS may have come from a lack of understanding of, and experience with, argument structure generally and Toulmin's concepts in particular, especially on the part of JH and KP; this also likely played a part in the inconsistencies in identifying CLAIMS (see below). BW who had the most experience in the study of argument found far fewer sentences in the final corpus to be EXTRANEOUS; she believed that many statements of external evidence and those making comparisons with previous work which JH and/or KP annotated as EXTRANEOUS

were in fact being used in support of a CLAIM, and thus were part of the authors' argumentation. Table 30 shows clearly this variation by annotator where BW identified only 54 sentences as EXTRANEOUS (0) but JH had 116 sentences in this category, and KP had 80. In the final corpus 101 of the 243 sentences which had two or three-way disagreements contained some variation between EXTRANEOUS (0) and another category. In Table 48 below I present the distribution of this variation by article and by conflicting category. Note that C6 is the only article with no such sentences. It also has the lowest percentage of EXTRANEOUS categorizations in the corpus (Table 29) at 5%, and is the only article with better inter-annotator agreement than under Model 1; both of these facts flow in part from the specifications for the WARRANT/BACKING (3) category (see Section 3.3.3.6), although elsewhere variation between WARRANT/BACKING (3) and EXTRANEOUS (0) was problematic (see below)

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Total
# Sentences	30	49	25	24	33	45	21	35	36	27	41	34	400
# with some Variation	23	24	20	22	22	14	8	23	22	15	23	27	243
# with Variation involving EXTRANEOUS (0)	15	8	14	9	9	0	5	8	6	4	12	11	101
Some (0) vs. (1) CLAIM	3	0	2	0	0	0	0	0	0	0	1	1	7
Some (0) vs. (2) GROUNDS	1	6	2	0	2	0	1	6	4	0	1	3	26
Some (0) vs. (3) WARRANT/BACKING	11	1	7	5	4	0	1	2	3	4	6	1	45
Some (0) vs. (4) QUALIFIER	4	1	2	3	3	0	3	0	0	0	5	8	29
Some (0) vs. (5) PROBLEM IN CONTEXT	0	0	3	3	1	0	0	0	0	0	1	0	8
Total 5 rows above	19	8	16	11	10	0	5	8	7	4	14	13	115

Table 48: Inter-annotator variation between EXTRANEOUS and other categories – Model 2

The category which conflicted most frequently with EXTRANEOUS (0) was WARRANT/BACKING (3); this variation occurred in 45 different sentences. Eleven of these were in article C1, where most of paragraph four contained descriptions of previous (generally cited) work, as discussed in Section 3.3.3.1. This variation is not surprising given that, especially for a non-expert, it is not always easy to see if external evidence and general facts are being given as background material or as support for a CLAIM in their current argument. Variation between EXTRANEOUS and QUALIFIER (4) was less frequent at 29 sentences; Section 3.3.3.12 gives examples found in article C12, which had the highest number (eight) of such sentences.

As discussed in 3.3.2.2, variation between EXTRANEOUS (0) and GROUNDS (2) can be the result of a non-expert in the biomedical content believing previous results to be current. It may also result from the specification of "Statements related to the methodology" for category (0) (Appendix D); this text was designed to keep 'background' to the current experiment out of the argument structure, but it is too broad to be appropriate. There is little variation with PROBLEM IN CONTEXT (5) but it is by far the least common category overall in the corpus at 6.3% (Table 29). Perhaps the most unexpected variation is between EXTRANEOUS and CLAIM (1); these are few in number and although some are 'legitimate' (see 3.3.3.1), at least one is the result of an error (in C11). In order to reduce this variation involving the EXTRANEOUS (0) category we need a) annotators to have a better understanding of argument structure, b) clearer specifications for this category, including amending the text re methodology, and possibly c) including EXTRANEOUS in the Trumping guidelines.

To examine the concept underlying the QUALIFIER category, we need Toulmin's view of qualification: for him it is related to notions of probability and commitment, similar to those in everyday speech where we use terms such as *possibly* and *probably*, both of which are hedges, as discussed in Section 1.4. In terms of argument:

> Our probability-terms come to serve, therefore, not only to qualify assertions, promises and evaluations themselves, but also as an indication of the strength of the backing which we have....By qualifying our conclusions and assertions in the ways we do, we authorise our hearers to put more or less faith in the assertions or conclusions, to bank on them, rely on them, treat them as correspondingly more or less trustworthy. (2003: 83-84)

In Jenicek's adaptation of Toulmin a qualifier provides the "strength, certainty or probability assigned to the claim" (2006: SR30) (see Section 1.3.2.2). But as adapted by Graves specifically to suit biomedical research texts, our category (4) QUALIFIER applies to statements of "possible explanations" for results or where authors "compare and contrast with external evidence" (Appendix D). This latter aspect represents Toulmin's idea above re "backing".

Unfortunately, however, this connection to external evidence led to considerable inter-annotator variation between the categories QUALIFIER (4) and WARRANT/BACKING (3): 46 sentences in the final corpus contained some variation between (3) and (4) (see Table 49 below). This variation was particularly pronounced in articles C4 and C5 as noted above in Sections 3.3.3.4/5. The source of this variation seemed to be the difficulty deciding whether the "compare and contrast with external evidence" aspect of QUALIFIER (4) or the straightforward "external evidence" of WARRANT/BACKING (3) was more appropriate for certain, especially complex, sentences; although (4) could Trump (3) if it made "external

evidence" relevant to their CLAIM, uncertainty often remained. (This issue is discussed further in Section 5.2.) More surprising, and more problematic given the importance of identifying CLAIMS, was the amount of inter-annotator variation between QUALIFIER and CLAIM (1): 52 sentences. The nature of articles C8 and C9, including the prevalence of hedging and the high number of CLAIMS (see 3.3.3.8/9), led to their having the highest number of (1) vs. (4) sentences, ten each. In C6, however, which also had a high number of CLAIMS we were more likely to be able to all agree on CLAIMS (Table 25) and hence had less (1) vs. (4) variation. Variation between QUALIFIER (4) and GROUNDS (2) generally occurred with statements relating current methodology to other studies, such as sentence 4-9 in article C11 (3.3.3.11). The least frequent category to conflict with QUALIFIER (4) was PROBLEM IN CONTEXT (5), with only half the articles having any such variation; again we note that category (5) was the least common across the corpus (Table 29). The distribution by article of variation between QUALIFIER (4) and other categories is presented below in Table 49:

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Total
# Sentences	30	49	25	24	33	45	21	35	36	27	41	34	400
# with some Variation	23	24	20	22	22	14	8	23	22	15	23	27	243
Some CLAIM (1) vs. QUALIFIER (4)	6	6	1	6	2	7	1	10	10	1	2	0	52
Some GROUNDS (2) vs. QUALIFIER (4)	2	1	3	0	1	2	0	3	0	4	6	4	26
Some WARRANT/BACKING (3) vs. QUALIFIER (4)	4	3	3	8	7	1	1	4	4	6	2	3	46
Some QUALIFIER (4) vs. PROBLEM IN CONTEXT (5)	1	0	1	5	1	0	0	0	0	0	2	5	15

Table 49: Inter-annotator variation between QUALIFIER and other categories – Model 2

Given the significance of the problem of variation between CLAIM (1) and other categories, but especially QUALIFIER (4), we return again to Toulmin to examine the core understanding of what constitutes a "claim": For him any "assertion" implies that one is making a "claim"; but from the point of view of argument, it is also a "conclusion". He stresses that a distinction must be made "between the *claim* or conclusion whose merits we are seeking to establish (C) and the facts we appeal to as a foundation for the claim – what I shall refer to as our *data* (D)." (2003: 90) This distinction makes sense from the point of view of logical structure i.e., "if D then C" (2003: 91); in natural language, however, and in our corpus in particular, these two components frequently occur in a single unit – the sentence. The hope was that the Trumping guideline, where CLAIM (1) Trumps categories (2) through (5) (Appendix D), would reduce inter-annotator variation in such complex sentences. Despite this, considerable variation involving CLAIM was found in the final corpus (see especially Table 49 above).

Some of Jenicek's specifications for a "Claim" are particular to clinical medicine, and do not apply to our data, but his general description is of a "proposition reached by reasoning; conclusion; solution to the problem in context". His "problem in context" is the starting point of an argument in medicine, including "Hypothesis; research question(s); Objectives". (2006: SR30) When I questioned him about this structure, moving from a statement of the problem to its solution, he replied: "Because philosophers dissect very often an argument starting by the claim, I have found [it] relevant to make a distinction between an 'initial' claim (your hypothesis or thesis) and the ('final') claim as defined by Toulmin." (personal communication 2007) This structure, designed to create

rather than analyze an argument (see Section 1.3.2.2), was adapted by Graves such that the argument structure does not begin and end with a Claim, but rather starts with a Claim, as a "proposition put forward based on analysis and interpretation of results" (as in Table 2). Under the revised Model 2 a statement of an open research question might in fact be considered as EXTRANEOUS (0) i.e., as background to the authors' argument.

A key problem with the definitions of the categories CLAIM (1) and QUALIFIER (4) in Model 2 is that frequently statements in our corpora seem to meet the specifications for both categories at the same time. For example sentence 4-4 of article C2 states: *We found that HDAC1 phosphorylation site mutants were significantly more sensitive to trypsin digestion compared to wild type HDAC1 (Figure 3), suggesting that HDAC1 trypsin sensitivity correlates with the interaction with associated proteins , enzymatic activity, or both.* This statement provides current results, followed by possible explanations for these results i.e., a QUALIFIER (4); but it is also a proposition based on "what the results mean" i.e., a CLAIM (1). This is a particularly common statement format in both our corpora, and more often than not it includes a hedging verb such as *suggest, may, could*, classic indicators of 'qualification' in Toulmin's original model. JH and KP annotated this sentence as CLAIM, and BW as QUALIFIER. They are correct in that they have followed the Trumping guideline, but BW clearly did not see 4-4 as a CLAIM. In order to assess when sentences of this type were CLAIM rather than QUALIFIER she looked at the overall argument structure, particularly as it was manifested in the preceding and following sentences, but also attempted to evaluate the rhetorical significance of the statement. The latter, however, likely requires a better understanding of the biomedical science and the

authors' audience than she had, and it may well be that JH and KP's annotations of CLAIM for 4-4 were correct. Given the extent of this (1) vs. (4) variation, and the pervasiveness of the type of sentence above in the *BMC*-series of journals, it is clear that the specifications for both categories, and perhaps the Trumping guidelines, need to be revised so that annotators will be better able to distinguish between them.

Although sentences with inter-annotator variation between CLAIM (1) and QUALIFIER (4) (in Table 49 above) were by far the most numerous and problematic in applying Model 2, there were conflicts between CLAIM and other categories; these are shown by article in Table 50 below (variation with EXTRANEOUS (0) is already shown in Table 48 above). An example of variation between CLAIM (1) and GROUNDS (2) occurred in sentence 1-7 of article C6: *This is the first report of GIRK protein in breast cancer cells.* This statement follows a series of six sentences where the authors present findings from their own previous studies (all agreed on WARRANT/BACKING (3) for these); BW annotated 1-7 as GROUNDS (2) but both JH and KP chose CLAIM (1). There seems to be no question that this is a rhetorically significant statement: being first is very important in scientific research generally, but in addition, to be addressing a human disease which affects millions of women makes it even more compelling. It is how they are introducing the 'marketing' of their current results. But is it a CLAIM? In the Toulminian terminology it appears to be an "assertion" but it is hard to see how it could be described as a "conclusion"; it seems more like a 'fact'. It is a proposition to which a truth-value can be assigned – other researchers in the field may in fact counter that the statement is false. But as stated in Appendix D a CLAIM should be "based on analysis and interpretation of

results"; this criterion is where BW believed that sentence 1-7 did not qualify as a CLAIM (1) and thus the Trumping guideline was not applicable. It is a statement about their study, rather than their results, but it is "internal evidence" and therefore she annotated it as GROUNDS (2). Nevertheless, it does seem to 'act like' a CLAIM, especially given the block of preceding evidence (1-1 through 1-6) where the authors are clearly presenting their experience and credentials, building to 1-7. Given that this type of sentence is not uncommon in these research texts, and that this is a CLAIMS-based Model, it is important to work with Graves, the author of the Model, on revising or expanding the specifications for this category.

Variation between the categories CLAIM (1) and WARRANT/BACKING (3) was relatively rare in the final corpus (see Table 50), as these two categories are generally clearly distinguished. The final sentence (5-6) of article C1 was the source of three-way disagreement: *The higher affinity of aromatase for 4-androstenedione than for testosterone seems to agree with the present data.* Neither this sentence, nor any of the previous five sentences in this paragraph, has any citation, but in 5-4 the authors refer to an interpretation of their current results being *in contrast with a generally believed pathway.* Based on this, and not knowing whether the phrasal subject in 5-6 refers to a known fact or a particular work, BW annotated 5-6 as QUALIFIER (4) – "compare and contrast with external evidence". KP categorized it as WARRANT/BACKING (3), but here this seems inappropriate as it misses the key connection (*agree with*) to their current results; also (4) could Trump (3) (Appendix D), and 5-6 does seem relevant to the

authors' CLAIM in 5-5 (we all agreed on CLAIM FOR 5-5). Although 5-6 seems more support for a CLAIM rather than a CLAIM itself, JH annotated it as CLAIM (1).

As shown in Table 29 article C12 has the highest number (nineteen out of a possible 102 tokens) of PROBLEM IN CONTEXT (5) sentence annotations in the final corpus. Thus it is not surprising that C12 accounts for half (five out of ten) of the sentences with variation between CLAIM (1) and PROBLEM IN CONTEXT (shaded) in Table 50 below. One such is the penultimate sentence in the Discussion section of C12: *Nevertheless, our study can serve as a guide for any kind of cartilage study, and reference genes could be used once tested for low M values.* (5-9) JH and BW annotated this as PROBLEM IN CONTEXT (5), implications of their current results for future research, but KP saw it as a CLAIM (1); it seems either category could legitimately apply. It is not known whether KP was conflicted between (1) and (5) and made use of the Trumping guideline, or simply believed it to be a CLAIM.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Total
# Sentences	30	49	25	24	33	45	21	35	36	27	41	34	400
# with some Variation	23	24	20	22	22	14	8	23	22	15	23	27	243
Some CLAIM (1) vs. GROUNDS (2)	1	2	0	0	0	2	0	2	2	1	1	0	11
Some CLAIM (1) vs. WARRANT/BACKING (3)	2	1	2	1	0	0	0	1	0	0	0	0	7
Some CLAIM (1) vs. PROBLEM IN CONTEXT (5)	0	1	1	1	1	0	1	0	0	0	0	5	10

Table 50: Inter-annotator variation between CLAIM (1) and other categories – Model 2

4.2.1.3 Annotators

Annotators are a source of variation as each person has their own unique personality and view of the world; these differences mean that each annotator will interact with the corpus data and the two Models of argument in their own individual way. It was believed that the lengthy training period, including consultation, feedback and practice with annotating, would allow all three annotators to 'be on the same page' i.e., to share an understanding of the goals of the project as well as the concepts behind, and the application of, both Models to the corpus data. Although I believe that this was the case, the fact remains that the results from the final corpus show considerable inter-annotator variation. It is impossible to completely isolate the variation caused by our individual differences, especially given that our inter-annotator agreement varied between Models and, widely, among articles (see Table 45 above). We have already looked at the breakdown of two-(and three-)way inter-annotator agreement sentences across the final corpus under both Models (Tables 22 and 23) and the overall distribution of argument category choices by individual annotators (Tables 27 and 30); here I compare each pair of annotators in terms of frequency of sentence categorizations on which they agreed and disagreed.

4.2.1.3.1 Pair-wise inter-annotator Crosstabulations – Model 1

In order to make these two-way comparisons I used SPSS to perform Crosstabulations for each pair of annotators under both Models 1 and 2. These Tables will be shown below; the cells on the diagonal (shaded) show the number of times that two annotators agreed on a particular argument category for a sentence. Again I note that these numbers will

reflect the overall category distributions for the Models (Tables 26 and 29) i.e., some categories occur more frequently than others. We also see the data on annotator predilections in Table 27 from a pair-wise agreement point of view.

The remaining cells show the distribution of categories where the two annotators disagreed. For example in Table 51 below in row 1 we see that JH and KP agreed on category (1) CONTEXT in 72 sentences; reading across row 1: for 8 sentences that JH annotated as CONTEXT (1) KP chose METHOD (2); 1 sentence JH annotated as (1) KP chose CURRENT RESULTS (3), etc. Reading down column 1 we see that there were 8 sentences KP annotated as CONTEXT (1) which JH annotated as METHOD (2) , 4 which JH annotated as CURRENT RESULTS (3), etc. I first present the three pair-wise distributions for Model 1 below:

	KP					JH Totals
JH	**1**	**2**	**3**	**4**	**5**	
1	72	8	1	0	11	92
2	8	25	4	2	4	43
3	4	5	48	3	7	67
4	29	2	5	13	8	57
5	11	6	5	3	116	141
KP Totals	124	46	63	21	146	400

Table 51: Annotator crosstabulation JH * KP – Model 1

	JH					BW Totals
BW	**1**	**2**	**3**	**4**	**5**	
1	72	11	4	20	14	121
2	4	26	3	3	3	39
3	3	3	48	3	2	59
4	3	0	6	22	5	36
5	10	3	6	9	117	145
JH Totals	92	43	67	57	141	400

Table 52: Annotator crosstabulation BW * JH – Model 1

	KP					BW Totals
BW	**1**	**2**	**3**	**4**	**5**	
1	93	13	2	3	10	121
2	5	28	2	3	1	39
3	4	3	48	2	2	59
4	11	0	9	12	4	36
5	11	2	2	1	129	145
KP Totals	124	46	63	21	146	400

Table 53: Annotator crosstabulation BW * KP – Model 1

I begin this discussion regarding Tables 51-53 by pointing out that since the overall three way agreement was at 60.5% under Model 1 we expect to see more pair-wise annotator agreement than under Model 2, where the three-way agreement was 39.3%. By totalling the shaded cells in Tables 51-53 we see that JH and KP agreed on a total of 274 sentences, BW and JH agreed on 285 sentences and BW and KP on 310 sentences (an average of 290). Somewhat surprisingly, given their similar academic experience, JH and KP were the most likely to disagree, while despite their different backgrounds, BW and KP were the least likely to disagree under Model 1. Although as referred to numerous times above there had been an expectation that based on their similar level of knowledge

in biomedical sciences JH and KP would be likely to agree on annotations, and to differ from BW who lacked a background in these fields, this was not the case under Model 1. In fact as shown in Table 22 the number of JK~B sentences was the smallest of the two-way agreement groups at 32; JB~K sentences numbered 42 and there were 69 BK~J sentences, almost as many as the first two groups together. Although it is true that BW and KP agreed on more sentences (77.5% of the 400 sentences) than the other two pairs, the spread among the three groups is only 36 sentences (274-310) or 9% of the 400 sentences. Thus it is striking to see how the data in Table 22, which include three-way agreement, do not give us the picture of how often two particular annotators agree, or disagree, that we see in Tables 51-53 (unlike under Model 2, see below).

It is also striking to see the parallels between Tables 51 and 52: when we compare the shaded cells we see that the numbers are virtually identical except for category (4); this latter is not surprising since JH, the common annotator in these two Tables, had a greater propensity for choosing RESULTS COMPARED (4) than either KP or BW (see Table 27). Of course the issue here is not that the pairs were agreeing on the *same* sentences – crosstabulations only give us frequency counts across the corpus – but it is interesting to see the common tendency to select particular categories under Model 1. Table 53 shows a distribution of agreed-upon categories similar to those in Tables 51 and 52, but with BW and KP agreeing on more category (1) CONTEXT sentences (93 vs. 72) and more ANALYSIS (5) sentences (129 vs. 116/117). This again reflects the fact that they were more inclined than JH to choose these categories (Table 27).

There were 44 sentences where JH chose category (4) RESULTS COMPARED and KP chose another category, the most common being CONTEXT (1) with 29 sentences (Table 51, row 4). The disagreements between the categories (1) CONTEXT and (5) ANALYSIS in Table 51 were reflections of each other: JH chose CONTEXT (1) and KP chose ANALYSIS (5) in 11 sentences, and there were 11 sentences where JH chose (5) and KP chose (1). In Table 52 the most pronounced variation occurred where BW chose category (1) CONTEXT: in 49 sentences JH chose a different category, including 20 sentences where he chose RESULTS COMPARED (4) and 14 sentences with ANALYSIS (5) (row 1). For ten sentences where BW chose ANALYSIS (5) JH chose CONTEXT (1), and RESULTS COMPARED (4) in 9 sentences (row 5). There was less variation between BW and KP since they agreed the most, but where BW selected CONTEXT (1) there were 13 sentences where KP chose METHOD (2) and 10 where she chose ANALYSIS (5) (row 1, Table 53) KP varied from BW on 24 sentences which BW annotated as RESULTS COMPARED (4), for 11 she had CONTEXT (1) and for 9 she had CURRENT RESULTS (3) (row 4). The variation between categories (1) and (5) in Table 53 was similar to that between JH and KP above (Table 51): KP chose CONTEXT (1) for 11 of the sentences where BW chose ANALYSIS (5) (column 1), and BW chose (1) for 10 of the sentences KP annotated as (5) (row 1).

4.2.1.3.2 Pair-wise inter-annotator Crosstabulations – Model 2

As noted above there was less pair-wise inter-annotator agreement (i.e., more variation) under Model 2 than Model 1: under Model 1 the average number of sentences with pair-wise agreement was 290 (see above) whereas under Model 2 it was 227 (see below). However unlike under Model 1, under Model 2 the number of sentences with agreement

in the three pair-wise groups in Tables 54 to 56 below is in parallel with the breakdown of two-way agreement/variation groups in Table 23: JH and KP, and also BW and KP, agreed on 228 sentences and BW and JH agreed on 225 sentences (for an average of 227); in Table 23 there were 71 JK~B, and BK~J, sentences, and 68 JB~K sentences. Tables 54 to 56 below present the crosstabulations for annotators and categories under Model 2:

	KP						JH Totals
JH	**0**	**1**	**2**	**3**	**4**	**5**	
0	65	7	5	29	8	2	116
1	0	44	2	1	2	0	49
2	2	3	51	0	5	0	61
3	4	4	2	32	7	0	49
4	8	27	10	18	25	10	98
5	1	6	1	5	3	11	27
KP Totals	80	91	71	85	50	23	400

Table 54: Annotator crosstabulations JH * KP – Model 2

	JH						BW Totals
BW	**0**	**1**	**2**	**3**	**4**	**5**	
0	51	0	1	1	1	0	54
1	1	26	1	1	11	5	45
2	18	5	50	3	9	1	86
3	24	0	0	31	25	1	81
4	16	17	9	13	50	3	108
5	6	1	0	0	2	17	26
JH Totals	116	49	61	49	98	27	400

Table 55: Annotator crosstabulations BW * JH – Model 2

	KP						BW Totals
BW	**0**	**1**	**2**	**3**	**4**	**5**	
0	31	1	4	14	3	1	54
1	2	37	2	1	2	1	45
2	18	7	55	2	4	0	86
3	13	2	2	56	3	5	81
4	11	39	8	9	37	4	108
5	5	5	0	3	1	12	26
KP Totals	80	91	71	85	50	23	400

Table 56: Annotator crosstabulations BW * KP – Model 2

Given that there was less agreement under Model 2, and that it has one more category than Model 1, it is not surprising that we do not see as many similarities across the Tables on agreement (the shaded cells) as we do under Model 1 above. The exception is category (2) GROUNDS, the category most often agreed upon across the corpus (Table 25), with 51, 50 and 55 sentences in Tables 54, 55 and 56 respectively. Below we discuss the most striking pair-wise variations; these are reflections of the variation by category for Model 2 discussed in 4.2.1.2 above, and the overall category choices by annotator shown in Table 30.

The key differences observed earlier in Table 30 are: KP's inclination to choose CLAIM (1) rather than QUALIFIER (4), JH's tendency to choose EXTRANEOUS (0) rather than WARRANT/BACKING (3) and BW's predilection to choose categories other than EXTRANEOUS (0); they are all seen here in more detail (the totals from Table 30 appear in the final rows and columns of Tables 54 to 56). In Table 54 it is interesting to see the distribution across categories of KP's choices where JH chose QUALIFIER (4): she varied

here on a total of 73 sentences, ranging from 8 sentences which she had annotated as EXTRANEOUS (0) to 27 which she had identified as CLAIMS (1) (row 4). Even though JH and BW have a similar total number of sentences annotated as QUALIFIER (4) at 98 and 108 respectively, Table 55 makes it clear that they agreed on only 50; of the 48 sentences where BW varied against JH's choice of QUALIFIER, she had annotated 25 as WARRANT/BACKING (3) and 11 as CLAIM (1) (column 4). We also see a wide distribution of conflicting categories where BW chose QUALIFIER (4): JH's choices spread across all five other categories, ranging from 3 (PROBLEM IN CONTEXT) to 17 (CLAIM) sentences for a total of 58 (row 4, Table 55), and KP ranged from 4 (PROBLEM IN CONTEXT) to 39 (CLAIM) sentences for a total of 71 (row 4, Table 56).

We are now able to see the variation in Table 48 between EXTRANEOUS (0) and other categories from the perspective of different annotators. Given that JH had by far the largest number of sentences annotated as EXTRANEOUS (116), we are interested in which categories the other two annotators selected in lieu of (0). In Table 54 (row 0) we see that KP overwhelmingly chose WARRANT/BACKING (3), but in Table 55 (column 0) BW's choices were more varied, although like KP, the category most frequently conflicting with (0) was WARRANT/BACKING (3). It is interesting to note that although BW and KP agreed on fewer EXTRANEOUS sentences than BW and JH – 31 vs. 51 – the balance of column 0 in Table 56 shows some similarity to column 0 in Table 55; the key difference is in the number of WARRANT/BACKING (3) sentences – 13 vs. 24, a reflection of the fact that KP has 36 fewer EXTRANEOUS sentences than JH.

Given that KP annotated 91 sentences as CLAIM (1), almost as many as BW and JH combined, we now look at the categories in conflict with her choice of (1). In Table 54 we see in column 1 that JH most frequently chose QUALIFIER (4) rather than CLAIM (27 sentences), and in Table 56, column 1, that BW also had QUALIFIER most frequently, with 39 sentences. The pair most likely to disagree between WARRANT/BACKING (3) and QUALIFIER (4) (which total 46 sentences in Table 49) was JH and BW with 38 (13 + 25) such sentences (Table 55); JH and KP varied between (3) and (4) on 25 (18 + 7) sentences (Table 54), and BW and KP on 22 (9 + 13) sentences (Table 56).

4.2.1.3.3 Annotator Errors

Given the ultimately subjective nature of the annotation tasks in this project, and the fact that we had no biomedical experts to whom we could compare our annotations, we use the term 'error' in a relative sense: We presume that 'errors' cover a continuum of annotator choices from what appear to be violations of category specifications to those that suggest an incomplete knowledge of argument and its structure. These may be the result of haste or inattention (i.e., 'performance errors'), lack of understanding of the Models and their categories, and/or their theoretical foundations, not enough time spent reading the entire article, insufficient training/practice, or any combination of these. Nevertheless, even with the extensive amount of time BW invested in gaining background for this project and performing all annotations, she still found some of her own annotations in the final corpus that in retrospect she would have done differently, including those she would describe as poor choices, and some as outright mistakes. Some examples of inappropriate annotation choices by all three annotators have been given in

Sections 3.3.3.1-12; here we provide other examples, some that reflect the problems with the Models seen in 4.2.1.1/2 above.

Under Model 1 a major source of inter-annotator variation was disagreement between the category CONTEXT (1) and other categories. Although in many cases it was extremely difficult to be sure what belonged in this category, some cases involved errors. BW annotated sentence 2-9 of article C5 as (1): *The analysis of the charges induced by both the EMS and Az-MNU treatment argued against the hypothesis that natural polymorphisms introduced by contamination of the Nipponbare population could be responsible for the observations.* Despite the lack of citation, the previous text (which did have multiple citations) plus their use of the past tense led BW to believe it described only previous results. Later, on seeing JH and KP's annotations, she realized this was an error: the authors were relating previous work to their current results, thus the correct annotation would be RESULTS COMPARED (4). Sentence 1-7 of article C7 states: *Worker bees are not sterile.* This is stated as a known fact, as background to the external study cited in 1-8; thus, despite having no citation, it clearly belongs to the CONTEXT (1) category, and KP's annotation of CURRENT RESULTS (3) is an error.

I have discussed earlier the clarification that category (5) under Model 2 (PROBLEM IN CONTEXT) refers only to the consequences of the authors' current study for the future of their field. JH's annotation of sentence 1-14 in article C5 as (5) is a clear violation of this specification: The authors report that *Nipponbare populations treated either with1.5% EMS or by sequential soaking in 1 mM sodium azide and 15 mN MNU, showed a similar*

density of putative mutations detected by TILLING, ~1/300 kb. This is simply a statement of a current finding i.e., GROUNDS (2) with no reference to the future. It seems likely that this was a 'performance error' on JH's part since under Model 1 he annotated this sentence as CURRENT RESULTS (3). In Section 3.3.3.8 I noted that article C8 contains numerous statements of speculation regarding possible explanations for the authors' results; in sentence 6-7 they cite an external study in support of one of these hypotheses: *Moreover, during differentiation of human NSCs, myc transgene expression was shown to be down-regulated [16]*. KP's annotation of this sentence as GROUNDS (2) is an error, as it is presenting external, not internal evidence. As in the example above, this seems to be in conflict with her Model 1 annotation for 6-7 which was CONTEXT (1), not CURRENT RESULTS (3).

As has been made clear in this thesis there was considerable inter-annotator variation in identifying CLAIMS (1) under Model 2, and an often wide range in the number of CLAIMS found in a particular article e.g., C2 (from four to thirteen) and C9 (from four to fourteen) (Table 31). Although some of this variation was the result of differences of opinion or scientific background, here I provide some examples of errors of judgement. The first paragraph of article C11 consists of background to their current study; sentence 1-2 states: *Thus, it has recently been shown that the so-called "door-to-needle" time is a critical factor in the survival of patients with sepsis [21]*. The work cited, from 2006, involves none of the authors of C11, and thus far in the article there has been no mention of the authors' own results; nevertheless, KP annotated 1-2 as CLAIM (1). This statement

is in no way a proposition based on analysis of their results, and does not qualify as a CLAIM.

Some of our variation involving CLAIMS appears to relate to our differing understanding of argument structure. Of the eight sentences in the final paragraph of article C8 we had three-way agreement on only one. This paragraph explores the possibility that the source of their results was a *third difference...the use of different proto-oncogenes for immortalization.* (6-1) i.e., their choice of methodology. Sentence 6-2 describes their current methodology, 6-3 and 6-4 provide external evidence regarding one of the two proto-oncogenes used, then 6-5 states: *v-Myc may be a more potent immortalization agent than c-myc.* Both JH and KP annotated 6-2 through 6-4 as EXTRANEOUS (0) whereas BW saw 6-2 as GROUNDS (2) and 6-3 and 6-4 as WARRANT/BACKING (3), that is, as not external to the authors' argument. Given her preceding annotations, KP's choice of CLAIM for 6-5 must be seen as an error: According to the specifications for Model 2 (Appendix D) only material that is "not directly related to a CLAIM" belongs in the EXTRANEOUS category, but 6-3 and 6-4 are external evidence leading toward the statement in 6-5; in other words, if 6-5 is a CLAIM (1), then 6-2 through 6-4 must be part of the argumentation leading to making it i.e., they cannot be EXTRANEOUS (0).

Also related to argument structure is the question of whether we have identified too few or too many CLAIMS in each Discussion section. As was noted above JH identified no CLAIMS in articles C10 and C12; although C10 had the lowest number of CLAIMS overall with two (see 3.3.3.10), BW and KP found a total of eight CLAIMS in C12 (Table 31). It

seemed to me that without at least one CLAIM there could be no argument, since Toulmin's core argument layout is "Given data D, one may take it that C [Claim]" or "D → So C" (2003: 91-92) (see 4.2.1.2 above). When I asked Graves her view on the absence of any CLAIM she replied: "Perhaps the paper is so badly written that [JH] couldn't figure out where the claim was implied...[No claim] at all seems unlikely. Every published paper should be making some kind of claim of new knowledge." (personal communication 2009) At the other extreme, in article C8 both JH and KP had a block of five CLAIMS in a row in sentences 5-6 through 5-10. All three annotators agreed that 5-6 was a CLAIM (*...a variety of stem cells exists each with its own committed limits of differentiation.*) and the next four sentences expand on that CLAIM, offering a variety of different scenarios as to how it might work. BW thus saw these all as QUALIFIER (4) – "possible explanations"; the authors use the hedges *may* three times and *would* four times. However, rather than explaining their results directly, the authors are giving possible explanations for the possible implication expressed in 5-6 (*...this may suggest that...*). JH and KP saw all five sentences as CLAIMS but BW believed that this was not a reasonable argument structure: five CLAIMS in a row would 'dilute' the strength of the argument. Nevertheless, JH and KP may be right, especially if 5-7 through 5-10 were seen as "Minor CLAIMS" (Appendix D). This type of argumentation is not uncommon in *BMC*-series articles and it is therefore important to make precise the correct approach to identifying CLAIMS in these argument structures.

4.2.1.3.4 Summary

From the above we can see that annotators are a source of variation when they have a 'bias' toward or away from a particular category, as can be seen in the distributions for Models 1 and 2 in Tables 27 and 30, or when they make errors in their annotations (the former may include the latter). With the exception of obvious errors, an annotator's bias does not necessarily mean that they are right or wrong, but that they vary in a particular category from the average percentage breakdowns shown in Tables 26 and 29. For example under Model 2, BW had only 54 sentences (or 13.5% of the total of 400) annotated as EXTRANEOUS (0); the average for the corpus for category (0) was 20.8%. JH on the other hand had 116 sentences (or 29.0%) annotated as EXTRANEOUS (0). These extremes meant that there was inter-annotator variation in the 65 sentences which JH categorized as (0) and BW annotated differently; the distribution of her five other category choices is seen in column 0 of Table 55 above.

Numerous other manifestations of similar inter-annotator variations can be seen in Tables 51 to 56 above; again I point out that there is more variation under Model 2 than Model 1. Given the longer range goal of training an automated system we need to reduce both the systematic (annotator bias) and random (annotator errors) inter-annotator variation; in other words, we need to have more consistency in performance among annotators. Looked at from the point of view of selecting annotators, one might then try to find 'similar' annotators; this is not easy, however, given the multiple aspects of skill, experience and personality that affect the annotation process. In this project the two annotators who were similar in age and academic experience (JH and KP) did not agree

with each other more than they did with BW, the 'dissimilar' annotator; in fact under Model 1 as discussed above KP was more inclined to agree with BW than JH (Tables 51 and 53).

The selection of annotators, however, cannot be fully addressed without taking into account the need to revise both Models of argument; we need to reduce the opportunities for annotators to disagree. In Sections 4.2.1.1/2 I have identified the major sources of variation in each of the Models and in Section 5.2 I will discuss possible revisions to the Models. For example, by amending the specifications for the categories CLAIM (1) and QUALIFIER (4) under Model 2 in order to clarify how to differentiate them, we could begin to reduce the extent of the (1) vs. (4) inter-annotator variation seen in Tables 54 to 56. In addition, as referred to in Section 4.2.1.3.3 above, annotators need to have a better understanding of argument structure, especially as it relates to Model 2 and the identification of CLAIMS. This could be achieved through better training, particularly to show in detail how arguments are built in example *BMC* articles before beginning training annotations. The issue of the 'ideal' annotator will be looked at in Section 5.3.

4.2.1.4 Corpus Data

As has become abundantly clear throughout this thesis, and as manifested in the wide range of inter-annotator agreement among articles in Tables 18 and 21, the corpus data can be a major source of inter-annotator variation (see Section 2.6). Although there is some degree of homogeneity in that both the training and final corpora come from the same series of *BMC* research journals, the articles varied in writing style and clarity,

argument structure, sentence length and complexity, level of technical jargon, etc. These variations in the data in turn interacted with the different approaches to rhetoric of the two Models being applied; for example, the article with the highest three-way agreement under Model 2 (C6) was ranked third lowest under Model 1 (Table 21). Sections 2.3.2.2.1-3 and 3.3.3.1-12 provide numerous examples of the relationship between the text of a particular article and its annotation results. Here I will summarize the major problems encountered when attempting to classify the corpus data by argument category.

Given that scientists build on their own previous work, as well as those of other researchers, it is frequently extremely difficult, especially for a non-expert, to establish whether a statement is referring to the authors' current experiment or their earlier studies, or both. Authors may not cite their own previous work, especially when they are also discussing a current finding, either within the same sentence or in a surrounding one. Uncited statements of what appear to be 'facts' i.e., generally accepted knowledge in the field can also be problematic for a non-expert; although sometimes their position in the article (e.g., the first paragraph is often background material) or the content of the surrounding text can offer clues, often it is difficult to decide whether they are referencing the current study or not. Under Model 1 this is usually a decision between CONTEXT (1) and another category and under Model 2 it is often between EXTRANEOUS (0) and another category. As discussed above, these two are both major sources of inter-annotator variation in this study.

The structure of an article can also lead to lower total agreement values. For example authors may choose to put most statements of their experimental results in the Results section of their paper rather than in the Discussion. Given that these types of statements are easy to agree on – CURRENT RESULTS (3) under Model 1 and GROUNDS (2) under Model 2 – their sparseness can lead to lower overall agreement. This can also lead to confusion as to whether sentences in the Discussion are referencing current or previous work.

The complexity of the argument structures which writers use may also affect the level of difficulty in the annotation task. Section 3.3.3.12 discusses the level of challenge posed by the researchers' results in article C12, and thus their need to carefully structure their argument in order for it to be accepted by their audience. Their opening with justification for their methodology led to disagreement, especially under Model 2, as to whether text was CONTEXT/EXTRANEOUS or was in fact part of the authors' argumentation. They open their final paragraph with: *Finally, it is necessary to take into account some limitations of our study.* (5-1) Eight of the nine remaining sentences expand on this: the authors are stating limitations, but at the same time saying 'we have followed accepted methodologies'. Within this overview of heading off potential criticism they discuss their method and results, as well as citing previous work. This led to inter-annotator variation under both Models, but mainly Model 2: How to decide if the key rhetorical purpose of a sentence was to give a limitation of their study (the overall point of the paragraph), or to discuss their methodology? The Model 2 annotations were all over the map with three-

way agreement on only the final sentence (5-10); they included KP identifying sentences 5-6 and 5-7 as EXTRANEOUS (0) but 5-3 and 5-9 as CLAIMS (1).

The above is also part of a larger problem - the effect of the surrounding text on the interpretation and annotation of a sentence. The final paragraph of article C8 (discussed in 3.3.3.8) is similar to that described above for C12: it explores the possibility that a third factor is responsible for their current results, with descriptions of their methodology as well as numerous citations of previous works. Thus although in isolation both sentences 6-6 and 6-7 are statements of external evidence, from the rhetorical perspective (and textually at the paragraph level) they are possible explanations for the authors' current results. These sentences produced two-way agreement under Model 1 but three-way disagreement under Model 2. In cases where it is not clear whether an uncited statement refers to the current experiment, previous work or a known fact, an annotator looks at the surrounding text for guidance; if both preceding and following sentences do have citations, it may be difficult to decide which is being referenced by the uncited sentence, especially for a non-expert. There is also the problem common to all discourse analysis – the scope ambiguity of discourse connectives such as *This*: To which aspects of the preceding text does deictic *This* refer?

Sentences which are grammatically and/or argumentatively complex are, by definition, difficult to categorize, under either or both Models (see for example sentence 6-8 in article C8, Section 3.3.3.8). I have addressed the issue of the 'ideal' unit of annotation elsewhere (White 2005b), but in this study practical considerations led to the decision to

use the sentence for articles T3-T5 and C1-C12 (see Section 2.3.1.4). As noted above the sentence length can vary tremendously both within an article and across the corpora, and longer sentences are more likely to be complex. This is also an aspect of writing style: some writers prefer to use more, shorter sentences while others combine multiple ideas into single sentences. The Trumping guidelines were developed to mitigate the conflicts which are inevitable when the sentence is enforced as unit of annotation; these proved to be useful in some situations, but they were certainly not entirely successful in reducing inter-annotator variation.

Finally there is the problem of technical content and lexicon in the *BMC*-series of journals. As with any academic specialty, even within a field such as Biochemistry there will be numerous sub-specialties, each with its own particular knowledge base and sub-language. Thus the ability to comfortably annotate the argumentation of a given biomedical article may involve a general level of scientific understanding and terminology, but also an awareness of a more specialized field. For an annotator who is not an expert in the technical content, more time and attention are required to read (and often re-read) the article; the task becomes even more difficult when the article is poorly (not clearly) written. Within this project's corpora there was a wide range of writing abilities on the part of the various authors, including some who appear to not be native speakers of English, from clear and elegant to muddled and confusing. It is presumed that this range of skills is representative of the *BMC*-series of journals in general and therefore that this is a variable that cannot be predicted or controlled.

4.3 Hedges

4.3.1 Frequency of hedges

As presented in Tables 19 and 33, there is a wide range in the use of hedging across articles in the two corpora; this inter-article variation may relate to preferences in structuring argumentation, writing style, degree of challenge to their audience posed by the authors' results, or a combination of these factors. Hedges are slightly more frequent in the training corpus – 78 in 147 sentences, or an average of one hedge every 1.9 sentences – than in the final corpus – 194 hedges in 400 sentences, or an average of one every 2.1 sentences. In both corpora verbs (both modal and lexical) are by far the most common grammatical hedge group: 79.5% of hedges in the training corpus and 80.9% in the final corpus.

Some hedges occur far more frequently than others across both corpora. The first and second most commonly occurring hedges in both the training and final corpora are *may* and *suggest*, and the fourth is *indicate*. The fifteen most frequently occurring hedges in the final corpus, and the twelve most frequent in the training corpus, are presented below in Table 57 in order from most to least frequent. Although there are thirteen hedges listed in Table 19, the verb *can* was deleted from the final set of hedges (see Section 2.5.3), thus it is not included below. For the purposes of comparison to studies external to our project, the most frequent hedges from Hyland's corpus of 26 research articles in Cell and Molecular Biology are included in Table 57 as well (1998: 149). His three hedges which were not considered in this project are shaded.

Order	TRAINING CORPUS	FINAL CORPUS	HYLAND'S RESEARCH ARTICLE CORPUS
1.	May	May	Indicate
2.	Suggest	Suggest	Would
3.	Possible/y	Could	May
4.	Indicate	Indicate	Suggest
5.	Likely	Would	Could
6.	Might*	Possible/y	About
7.	Could*	(Un)likely	Appear
8.	Would*	Should	Might
9.	Appear**	Seem*	Likely
10.	Seem**	Believe*	Propose
11.	Perhaps***	Possibility	Probably
12.	Think***	Appear**	Apparently
13.		Might**	Should
14.		Assume***	Seem
15.		Assumption***	Possible

Table 57: Order of most frequently occurring hedges in three corpora (*, **, *** indicate ties)

The similarities between the most frequently used hedges in the two corpora for this project and that of Hyland in Table 57 are striking, although there is variation in their orderings. We note that the order of hedges in the 11 to 15 positions in our corpora is of less interest than those more frequently occurring as the former have few instances e.g., *think* occurs only once in the training corpus and *assume* only twice in the final corpus. As noted in Section 1.4.1.3 adjectives such as *about* were not included in this study. The verb *propose* and the adverb *apparently* were not considered as the list in Table 3 was developed based on BW's experience with the Discussion sections of *BMC* articles and the hedges she most commonly encountered in them.

I point out three other differences between our study and that of Hyland. He states that the "rhetorical distribution [of hedges] follows expected patterns for pragmatic devices, with 84% occurring in the Results and Discussion sections of the RAs." (1996: 259), but he collected his hedging data from entire Research Articles (RAs), not only the Discussion sections. Also, his corpus is larger in number of articles, 26 vs. our five and twelve, and comes from a single field, Cell and Molecular Biology. The fact that, for example, our most frequent hedges are *may* and *suggest* whereas he has *indicate* and *would*, may be explained by the fact that his corpus is composed only of articles from this one field, whereas our corpora of biomedical research articles come from a wide array of different fields. Despite these differences the most notable comparison in Table 57 is the fact that the five most common hedges in our final corpus are identical to the top five in Hyland's corpus; this suggests some degree of commonality across biomedical fields in academic writing.

4.3.2 Analysis of hedging

In terms of lexical hedging verbs there is some degree of subjectivity in the interpretation of the 'strength' of a hedge. Perhaps the strongest commitment to a statement would come from 'prove' (e.g., *we have proven that*), with 'show' (which occurred 21 times in the final corpus) seeming somewhat weaker (e.g., *we have shown that*); moving into hedging, *indicate* (e.g., *our results indicate that*) seems a stronger commitment than *suggest* (e.g., *our results suggest that*). JH, however, considered *suggest* and *indicate* as "interchangeable". Hyland seems to put *believe* and *suggest* on equal footing: "A non-factive predicator, such as 'believe' or 'suggest' commits the speaker to neither the truth

nor falsity of a proposition" (1998: 44). Our results in Table 57 are consistent with Hyland's corpus where "A comparison with the other academic corpora shows that...particularly 'indicate' and 'suggest' are more prominent in scientific writing while 'seem' and 'assume' occur far less often." (1998: 126)

Regarding epistemic modal verbs, although in some contexts *might* and *may* could seem equivalent, Hyland states that *might* "expresses a higher degree of conditionality or tentativeness" than *may*. (1998: 117) In both the training and final corpora for this study, *may* is the most frequently employed hedge, with *might* well down the list; in Hyland's corpus *may* is number three and *might* is number eight (Table 57). When comparing his scientific corpus to other non-scientific corpora which he examined Hyland states: "A broad generalization is that while 'might' appears to occur slightly less frequently in scientific writing, 'may' occurs more often." (1998: 116) The fact that *may* occurs far more often than *might* in the *BMC*-series of articles used in this study could result from authors not wanting to be overly "tentative", or it could be simply that *may* is the more 'standard' means of expressing propositions in the *BMC*-series, or both. In fact Hyland states that the frequent use of hedges such as *indicate, suggest, could* and *should* by the writers in his science research articles "may result from conventionalism within the discourse community as a result of readers' constant exposure to them." (1998: 148) This would be consistent with any genre of academic writing where adopting the standard style and lexicon is a way of having your work accepted and published in your field.

4.3.3 Hedges and argument categories

As noted in Section 3.3.4, under Model 1 there was more three-way inter-annotator agreement on the categorization of sentences containing hedges than there was on all sentences in the final corpus: 73.7% vs. 60.5% (Tables 36 and 42). In Table 38 we see that almost three quarters of hedged sentences were annotated as ANALYSIS (5); this seems expected given that this category is where authors present speculation and probabilities: possible explanations, interpretations, significance, future directions, etc. I also note that although ANALYSIS (5) accounted for 36.0% of all annotations in the final corpus, 44.6% of sentences with three-way agreement were in category (5) (Table 28). But even though ANALYSIS is the most frequent category in the final corpus, and the most readily agreed upon, the fact that overall inter-annotator agreement is 13.2% higher for hedged sentences is still a striking result. It also suggests that the presence of lexical hedges might be a useful cue in predicting that a sentence is of the ANALYSIS category.

The next most common categorization for hedged sentences under Model 1 was CONTEXT (1) at 14.8% (Table 38). One group of such sentences are those where previous authors were hedging their findings in the past e.g., sentence 1-5 in article C10 (with verbs underlined to highlight the double use of the past tense, and hedges shaded): *Andrikovics et al recently reported a higher frequency of R219K in controls than in Hungarian stroke patients and [redacted] a protective role for this polymorphism [25].* Current authors may address an open research question in the present (see the final verb *is*) about an on-going "possibility" flowing from past work e.g., sentence 7-1 in article C2 (again verbs are underlined): *Previous work found that binding with Sin3, RbAp48, MTA2 and CoREST*

was lost whether HDAC1 was singly or doubly mutated at S421 and S423, conjuring the possibility that phosphorylation at either site is functionally redundant in vivo [24]. In both of these examples the presence of a citation, together with the use of the past tense suggests that CONTEXT (1) is the most likely categorization. Another example of a hedged sentence in the CONTEXT category is 1-1 of article C6: *Our previous work has indicated that breast cancer cells express mRNA for the GIRK channels.* Here there is no citation but the adjective "previous" along with the lexical hedging verb being in the present perfect tense make an annotation of CONTEXT likely. All three of these example sentences had three-way inter-annotator agreement on the category CONTEXT (1) under Model 1.

Under Model 2 there was less three-way inter-annotator agreement on the categorization of sentences containing hedges than there was on all 400 sentences in the final corpus: 35.1% vs. 39.3% (Tables 37 and 43). More than half (56.7%) of these categorizations involved sentences with two-way inter-annotator agreement, slightly more than the 52.5% for the overall annotations. Recall that disagreement between the categories of CLAIM (1) and QUALIFIER (4) was a major source of inter-annotator variation in the final corpus (see Table 49), and note that these two categories together account for approximately two-thirds (67.0%) of the hedged sentence classifications (Table 39). This is in marked contrast to the Model 2 category distributions shown in Table 32 where the combined percentages of CLAIM (1) plus QUALIFIER (4) are 36.7% for all annotations and only 29.9% of sentences with three-way agreement. Thus the presence of a hedge in a sentence is a strong indicator of either the CLAIM or QUALIFIER category. As already

discussed in detail in Section 4.2.1.2, it is important that these two categories become more clearly differentiated before future applications of Model 2.

The next most frequent categorization for hedged sentences under Model 2 is EXTRANEOUS (0) at 11.9% (Table 39). I note that the hedges found in EXTRANEOUS sentences are predominantly modal verbs (46 of 69 hedge occurrences or 66.7%); this is in contrast to QUALIFIER (4) where modals account for 51.2% of hedges and CLAIM (1) where they are only 36.1% of hedges. (It also contrasts with the CONTEXT (1) category under Model 1 where modals account for slightly less than half (48.8%) of hedges.) The EXTRANEOUS category was such a source of inter-annotator variation in the final corpus (see Table 48) that Model 2 will need to be revised to amend or eliminate this category; thus it is premature to speculate regarding the relationship between hedges and this Model 2 argument category.

4.4 Argument Type

During the training phase of this project I had a number of lengthy discussions with JH and KP regarding their experience with both scientific writing and experimentation, especially focussing on the different aspects and kinds of scientific discovery. Even with this investment, and the development of a new set of Argument Types (Table 20), we did not all agree on Argument Type for a single article in the final corpus (Table 41). It seems that rather than being a useful 'macro' guide to selecting micro-level argument categories (under either Model of argumentation), the identification of Argument Type was an exercise that seemed essentially unrelated to the annotation process. As discussed

in Section 3.3.5 the complexity of content and unique argumentation in each corpus article made it difficult, if not impossible, to decide definitively on a single Type. The four Argument Types in Table 20 may be too general, and have too much overlap between them, to be meaningfully applied to the *BMC* articles.

CHAPTER 5 CONCLUSIONS

5.0 Introduction

The primary goal of this study was to evaluate two Models of argument by applying them to a corpus of biomedical research texts; with the longer range goal of developing automated tools for Information Extraction, the aim is to develop Models which can be applied with minimal inter-annotator variation. The secondary goal was to investigate the performance of different annotators by having a lengthy training process, including feedback and discussions, as well as detailed analyses of the results by annotator for the final corpus. Ultimately, for annotated data to be reliable and useful for researchers there is a need for a Model of argument that is relatively easy to understand and apply, matched with annotators who are comfortable with the corpus content and familiar with the concepts of argument and its structure. This, of course, is far more difficult to achieve than to describe. What this study has shown is that complexity is the rule rather than the exception: the corpus data have a wide range of writing and argument styles, as well as technical sub-languages, and often sentences did not fit easily into single Model categories; annotators varied in their understanding of how authors were arguing and in their interpretation of the argument categories.

This thesis has already looked at a number of sources of inter-annotator variation, focussing especially on problems with the two Models of argument. In this Chapter the current results are situated in the context of other approaches, and the issues of reliability and how to evaluate it are addressed (Section 5.1). Suggestions for revisions to the Models, and for improving annotation methods and protocols, are presented (Sections

5.2-5.4). In addition future research directions, including the possibility of using hedges as cues to rhetorical category are discussed (Sections 5.5-5.7).

5.1 Evaluating Agreement and Reliability

In the field of Content Analysis, Krippendorff stresses that in designing studies and evaluating their results, analysts must take into account the 'reliability' of their data; they must safeguard "against the contamination of scientific data by effects that are extraneous to the aims of observation, measurement and analysis." (1980: 129) The key issue in achieving reliability is not the coders, but the process: "data should at least be reproducible by independent researchers, at different locations and at different times, using the same instructions for coding the same set of data...[and coders must work independently] lack of independence is likely to make data appear more reliable than they are." (1980: 132) He also notes that the choice of threshold for validity should depend on what one intends to do with the data: "Where possible, standards for data reliability should not be adopted ad hoc. They must be related to the validity requirements imposed upon research results, specifically to the costs of drawing wrong conclusions." (1980: 147)

From a statistical point of view, Krippendorff points out that measures of correlation and association are inadequate to assess reliability since they do not take into account errors made by coders, including "intra-observer inconsistencies" made over time. (1980: 130) He developed the alpha coefficient in order to evaluate inter-coder agreement by correcting for chance agreement. In 1996 Carletta strongly advised that the CL field

should adopt such a measure from Content Analysis in order to evaluate results from studies involving subjective human judgements such as those in discourse and dialogue tagging. She recommended the Kappa Statistic, a variant of Siegel and Castellan's kappa (1988), a coefficient "closely related" to Krippendorff's alpha. (1996: 252). Carletta's Kappa measures pair-wise agreement between coders making category judgements, correcting for the proportion of expected chance agreements. Where there are no prior data on which to base 'expected' values, they must be estimated from current annotation results; 'chance agreement' is thus "the agreement expected on the basis of a single distribution which reflects the combined judgements of all coders" (Artstein and Poesio: 564) In this study, the closest approximations to this for each of our Models are found in Tables 26 and 29, where we see the overall distribution of categories by all annotators; but as shown below, there are problems with Kappa.

Since that time the Kappa Statistic has become the de facto standard in evaluating inter-coder agreement in CL studies involving human judgements. More recently, however, questions have been raised in CL about the appropriateness of applying Kappa across the board in corpus annotation studies (Di Eugenio and Glass 2004, Craggs and McGee Wood 2005, Reidsma and Carletta 2008, Artstein and Poesio 2008, among others), and there is ongoing debate in the field of Content Analysis as well (e.g., Krippendorff 2004). In particular the calculation of the Kappa Statistic where there are more than two annotators may not be appropriate, and the assumption of annotator independence cannot always hold in CL studies. Numerous other coefficients are being discussed e.g., Cohen's

kappa[6], Scott's pi, Fleiss's multi-pi and weighted alpha. Questions that are being posed include: Which coefficient is appropriate to apply for a particular task? Or for particular types of data i.e., nominal, ordinal, interval, ratio? Have its underlying assumptions been met? Does it work for more than two annotators? If so, how does it calculate the measurement(s)? Are the units of annotation chosen reasonable for purposes of comparison? What is the appropriate agreement threshold for a particular data usage? How should one interpret the various coefficients?

In particular Di Eugenio and Glass discuss the use of Kappa in validating coding schemes – the core goal of the current study – where a 'good' value for Kappa means that the "categories are 'real'." (2004: 98) They point out that there are two "unpleasant behaviours of Kappa" (2004: 98) that are highly problematic for studies such as ours, having to do with how the probability of agreement by chance is calculated. Cohen's Kappa "is calculated from each coder's individual probabilities" (2004: 99) and thus is affected by annotator bias, such as has been shown in Tables 27, 30 and 51-56. On the other hand, Siegel and Castellan's Kappa assumes an equal distribution of categories (the "prevalence problem"), which is not the case here (Tables 26 and 29), and rarely holds in any discourse or dialogue-tagging studies. (2004: 100) In addition, this latter assumption "masks the exact source of disagreement among the coders. Thus, such an assumption is detrimental if such systematic disagreements are to be used to improve the coding scheme (Wiebe, Bruce and O'Hara 1999)." (2004: 96) As has been shown repeatedly in

[6] CL researchers report on numerous inconsistencies in the literature regarding terminology with agreement coefficients (Artstein and Poesio 2008) and other confusing issues such as there being multiple versions of the kappa coefficient (Carletta 1996).

this thesis, both the above assumptions are violated in the results of this study; therefore, rather than trying to apply a coefficient such as Kappa, the goal of this study has been to 'unmask' as many sources of disagreement as possible, and thus improve both Models of argument.

In addition to the complexity of trying to choose an appropriate coefficient of agreement for CL studies, Artstein and Poesio also stress the difficulty in interpreting the resulting values:

> We view the lack of consensus on how to interpret the values of agreement coefficients as a serious problem with current practice in reliability testing, and as one of the main reasons for the reluctance of many in CL to embark on reliability studies. Unlike significance values which report a probability (that an observed event is due to chance), agreement coefficients report a magnitude, and it is less clear how to interpret such magnitudes. (2008: 591)

In fact in discussing Kappa-like coefficients they state: "deciding what counts as an adequate level of agreement for a specific purpose is still little more than a black art" (2008: 576). And Craggs and McGee Wood reflect this view as well: "there are no magic thresholds that, once crossed, entitle us to claim that a coding scheme is reliable. One must decide for oneself, based on the intended use of a scheme." (2005: 294) The detailed analyses of the multiple factors and dimensions in the variation identified in these results will serve as input to decisions on appropriate agreement coefficients for future studies (see 5.7 below).

An excellent survey of a range of agreement coefficients which could be considered for CL annotation studies is found in Artstein and Poesio (2008). The details of the mathematics and assumptions behind these coefficients are beyond the scope of this thesis, but here I wish to focus on two key points found in this and other recent articles: a) Rather than simply quantifying how much disagreement exists, it is crucial to examine its sources (Is it the coding scheme? The annotators?) and its form (Is there systematic variation? Or random errors?) and b) the goals of Content Analysis – to look for correlations among different variables – are not necessarily the same as those in CL, where we want human-annotated data to train automated IE systems. These topics are addressed below.

5.1.1 Inter-annotator Variation in Current Results

As has been amply demonstrated in this thesis the results of the current annotation project show a range of systematic inter-annotator variation, such as an annotator's bias toward a particular category (Tables 27 and 30), differences between the distributions of argument categories where all three annotators agreed and overall category distributions (Tables 28 and 32), and conflicts between argument categories that are not clearly differentiated (Tables 46-50); in addition, there are noise-like (error) disagreements. Although Model 1's performance was better than that of Model 2 based on average overall inter-annotator agreement statistics, the levels of inter-annotator variation are high enough to warrant revisions to the Models, with the goal of reducing the systematic variation seen in Chapter 4 as much as possible. The most surprising result was the wide range, under both Models, of agreement among the articles in the final corpus (Table 21); although

undoubtedly a reflection of all sources of variation (see below), this also brings up the question of whether annotators should be chosen based on their expertise in a particular domain (see 5.3 below).

Below is a summary of the possible sources of inter-annotator variation that have been found in this study's corpora and discussed in this thesis:

- not understanding/misunderstanding the content of an article (science (sub)field)
- not understanding the Model/category
- variation in interpretation of the Model/category
- combination of the above
- not enough experience with annotation task/understanding argumentation
- not taking sufficient time to evaluate the content/argumentation/categorization
- individual annotator differences of opinion (subjectivity)
- not understanding/misinterpreting the authors' argumentative strategies (at macro or micro (sentence) level)
- variation in surrounding text e.g., choice of category for a sentence may lead to variation in the previous or following sentence

Given that this was a study designed to compare and evaluate Models of argument, rather than to produce data 'reliable' enough to be useful for training automated systems, the expectation was that these preliminary results would serve as diagnostics for identifying problems with the Models. Analyses of results throughout this thesis show a number of

problems with both Models, particularly in terms of specifications for certain categories (see 4.2.1), which will be addressed in Section 5.2 below. Also, I was not concerned with Krippendorff's issue of reproducibility over multiple annotations with different coders, but rather with examining the results from specific coders, and comparing their annotations, in order to examine variation stemming from individual differences. As Craggs and McGee Wood point out: "there is no need to calculate kappa in order to observe [annotator] bias, since it will be evident in a contingency table of the data in question." (2008: 292) Variation between and among annotators has been seen in Tables 27, 30 and in the crosstabulations in Sections 4.2.1.3.1/2.

It is also worth noting again that there will always be some degree of inter-annotator variation in tasks such as argument analysis where subjective judgements are involved; there are no unequivocally 'right' answers as there are in, for example, part-of-speech tagging. Craggs and McGee Wood define subjectivity as "the absence of an obvious mapping for each unit of analysis onto categories that describe the phenomenon in question." (2008: 293) They note that there is concern regarding research in discourse and dialogue

> that the subjectivity of the phenomena being coded may mean that we never obtain the necessary agreement levels [to achieve reliability]...However, the fact that we consider these subjective phenomena worthy of study shows that we are, in fact, 'willing to rely on imperfect data', which is fine as long as we recognize the limitations of a scheme that delivers less-than-ideal levels of reliability and use the resulting corpora accordingly. (2008: 293)

It is clear from the annotation problems presented in detail in Chapters 2 and 3 that both Models of argument applied in this project require revisions before being used in further studies. In future work, this, along with considerations regarding the choice of annotators for rhetorical analysis (see Section 5.3 below), are necessary before being able to evaluate how much of the inter-annotator variation comes from such inter-personal differences. It has not yet been determined what level of such subjective variation would be acceptable in training data for argument analysis in IE. Whatever model of rhetorical analysis is applied to biomedical research corpora, some degree of subjective inter-annotator variation is inevitable. In a recent study Andrews et al. used experts from three different professional coding services to code concepts, and degree of certainty, found in data on clinical research using SNOMED CT (Systematized Nomenclature of Medicine, Clinical Terms); their results showed no significant level of agreement among the experts in any area, and there was three-way agreement on core concept only 33% of the time (2007: 501). The fact that even professional coders exhibit such a high degree of disagreement suggests that, even with revisions, the two Models of argument applied here may not achieve high levels of inter-annotator agreement; acceptable thresholds will need to be established in the future, within the context of particular applications (c.f. Craggs and McGee Wood 2008 above).

5.1.2 Computational Linguistics vs. Content Analysis

Despite the generalized adoption in the CL field of reliability measures from Content Analysis, several recent articles in the literature have warned that this should be re-examined, especially in light of the growing number of corpus annotation projects and the

increasing complexity and sophistication of coding schemes being applied (Artstein and Poesio 2008, Reidsma and Carletta 2008). Generally in CL, annotated data are used to train automated IE tools; this is a very different goal from that of hypothesis-testing, common in Content Analysis. Specifically, automatic classifiers learn by finding, and thus predicting, patterns in data. If the variation found in a hand-annotated corpus is a blend of systematic patterns and errors, as is the case with the results of this project, then the data are not appropriate for Machine Learning: "systematic disagreement [is] dangerous, because it provides an unwanted pattern for the learner to detect." (Reidsma and Carletta: 320) In other words, the machine reproduces the inconsistencies produced by human annotators. It thus becomes crucial to identify, as we have done in this study, patterns in variation such as annotator bias, confusion between categories CLAIM (1) and QUALIFIER (4) under Model 2, etc. It is also apparent that different dimensions of the variation may interact with each other; for example, an annotator with particular expertise in the biomedical content of a corpus article may be better able to evaluate whether a sentence is CONTEXT (1) or not under Model 1. In order to reduce the systematic variation which we have identified in this thesis, I make suggestions below regarding the Models of argument, the selection of annotators, and the annotation process.

5.2. Models of argument

I preface this Section by stressing that no revisions should be made to either Model before consultation with biomedical researchers – the 'end users' of the tools this work will help to create. Although previous studies have used 'experts' in biological sciences as annotators (Lu et al. 2006, Kim et al. 2008, Watters et al. 2005), I am not aware of any

model of rhetorical analysis that has been developed by, or in consultation with, biomedical researchers. Previous to this project I interviewed biomedical researchers and graduate students to ask them how they did research e.g., how did they select articles to read, what did they read first, which sections were most important, etc. The next step is to ask researchers if the Model categories we have been using are appropriate – do they differentiate classes of data that are meaningful and/or useful to them? For example, there were major problems with variation under Model 1 with the CONTEXT (1) category; is the idea of trying to separate 'old' from 'new' information important to them? Or are there other dimensions in biomedical data that should be brought into the Model? Although there was less inter-annotator variation under Model 1 than Model 2, it may be that the latter, based more on argument structure, could be revised and developed to be more useful to biomedical researchers than the former. Would researchers prefer more complex or simpler versions of the Models we have applied here?

As mentioned above, and seen in Table 46, the CONTEXT (1) category was a major source of variation under Model 1. It may be that the specifications are just too broad, and perhaps a return to a category such as the original 'Previous Work/Undisputed Facts' (Table 1) would be more appropriate. The fact that 'old' and 'new' material are so intertwined in biomedical literature was also reflected in the variation between CURRENT RESULTS (3) and RESULTS COMPARED (4); biomedical researchers may be able to help with the question of whether (3) should Trump (4), or the reverse. Allowing units of annotation at the clause, rather than sentence, level could help to reduce some of the above variation (see below). I would also ask researchers if keeping a specific category

for METHODS (2) in the Discussion section is useful, given that they have access to details of methodology in the Methods section. Possibly the CURRENT RESULTS (3) category could be expanded to a CURRENT STUDY category which might include text related to methodology.

Under Model 2 the category EXTRANEOUS (0) was involved in considerable inter-annotator variation (see Table 48). This problem reflects the fact that old and new information are frequently merged (e.g., the choice of current methodology is based on a previous study), but also the difficulty for annotators in identifying argument structure. If an annotator is unable to see how the authors are developing an argument across a text, they will not be able to recognize what is external to that line of argument. (Even with annotators relatively skilled in argument analysis, however, subjective differences will still exist.) Allowing units of annotation smaller than the sentence could help to alleviate some of the above variation, although this is not without risk (see below). The other major source of inter-annotator variation under Model 2 was the category QUALIFIER (4). As shown in Table 49 there was variation between (4) and other categories, but the most problematic variation is that between QUALIFIER (4) and CLAIM (1), given that CLAIM is the core argument category in Model 2. As discussed in Section 4.2.1.2 it was extremely common to find sentences that were at the same time both a "proposition based on analysis of results" and a "possible explanation for results". The specifications for these two categories should be amended to take into account these types of sentences such that the distinction between them is more readily apparent to annotators. Once again, a familiarity with argument structure would also be helpful. Given the amount of variation

between QUALIFIER (4) and WARRANT/BACKING (3), it is worth considering separating out the "compare and contrast with external evidence" aspect of category (4), perhaps in conjunction with revisions to the CLAIM (1) category. Any revisions to Model 2 will take place in collaboration with Graves, the developer of the Model, following consultation with biomedical researchers.

As mentioned in Chapter 2, I saw the use of the sentence as unit of annotation as a compromise: during training annotations of articles T1 and T2 it seemed that the added complexity of allowing units smaller than the sentence, especially given that we did not necessarily agree on what they were, risked compromising the focus on the Models being applied. Allowing annotators to split sentences into segments creates another layer of variation before the Models are even applied, and it also produces more units on which to disagree. The sentence does provide a readily identifiable unit for automatic analysis and extraction, but it is problematic in the case of complex sentences. This is especially true when looking at argument analysis where, as we have seen in both corpora, sentences containing segments from different categories, under both Models, are quite common. As stated earlier, and in general agreement with Mizuta et al. (2006), I believe the clause is the most appropriate unit of argumentation; unlike Mizuta et al., however, I would only allow clauses with a tensed (finite) verb. This segmentation into clauses would be a preprocessing step, done either by humans or machine; all annotators would then be working with the same units.

This use of sub-sentential units of annotation would simplify the Models by reducing the need for Trumping guidelines, but at the same time, as more units are created, there would be more opportunities for inter-annotator variation. The notion that Trumping was to be used where annotators were uncertain, e.g., believing more than one category might apply to a given sentence, led to an unanticipated problem: in some cases one annotator was not conflicted, but should have been, so did not use the Trump system, while another did apply Trumping. These types of inconsistencies may have created variation rather than reducing it. I still believe a Trumping system is useful, and more appropriate and flexible than a binary decision-tree approach. It may, however, be necessary to 'force' Trumping: if a unit contains material from more than one category, category x must Trump category y. Of course inconsistencies and variation will not be eliminated with this approach, but they should be reduced. It also seems worthwhile to have annotators record both the category they select and the category they eliminate by Trumping; these data would prove useful in future evaluations of both the Model being applied as well as its Trumping structure and utility. If time allows, having annotators report the basis for their decision would be even more informative.

5.3 Annotators

It would seem that the 'ideal' annotators for biomedical research texts are those that have training in both medical sciences and rhetorical theory; however, the probability of finding such individuals in a given research setting, never mind in a group large enough to provide sufficient inter-annotator data, is extremely low. The more likely scenario is a version of what we have in this annotation project: one person with knowledge of rhetoric

and linguistics, and two people with (senior undergraduate) knowledge of medical sciences. However, in my case I also had experience with annotating biomedical texts under both the ZA and AZ models (see Section 1.2) (White 2005a); I believe this has made me a 'better' annotator, or at least one able to train others. On the other hand, given my high level of investment in this project, it may make me tend to over-analyze the data, and not rely sufficiently on my intuitions.

At the fourth-year undergraduate level my annotators have not yet amassed the depth of biomedical knowledge that senior graduate students or researchers working in the field have. They are, however, far more familiar than I with biomedical terminology and methodologies, and generally better able than I to recognize the significance of a finding in an article; in terms of Model 2, for example, they might be able to see that a sentence I annotated as GROUNDS (2) is actually a CLAIM (1). As Graves aptly described them, they are not yet experts, but rather "quasi-informed". Since they both applied for, and accepted, the position of "Biomedical text annotator", I assumed they had more interest in language and/or writing than the average fourth-year medical science student. Despite the fact that these students were not highly paid, and were performing this annotation work while carrying a full course-load, I believe that their commitment to this project supported my hypothesis. Their degree of engagement was a relevant factor in the value I place on the results from this project.

Thus the current project involved annotators who were neither 'expert' nor 'naïve', but somewhere in between. Although there is some truth to the notion that there are no

'experts' when it comes to subjective judgements (Carletta 1996, Teufel 1999), there is no question that annotation of biomedical texts such as those in this study should be done by those with some knowledge of the domain being addressed. An annotator's scientific knowledge should be at least at a graduate level in order to be familiar with the technical content. The ideal annotator would also have some knowledge of rhetoric and argumentation, but annotators with expertise in both of these areas are likely to be extremely rare. The more realistic approach may be to involve two sets of annotators, one with expertise in argument and its structure and one with biomedical domain knowledge; these annotators could work together in pairs (see 5.4 below). In either case they should have a good understanding of English, as well as knowledge of and experience in academic writing.

Unlike in Content Analysis where reproducibility and annotator independence are required, CL projects frequently involve discussion and collaboration among annotators, and annotation schemes and guidelines may be revised over the course of a project (e.g., Lu et al. 2006, Kim et al. 2008).

> [I]n CL, corpora constitute a resource which is used by other processes, so the emphasis is more towards usefulness. There is also a trade-off between the sophistication of judgements and the availability of coders who can make such judgements. Consequently, annotation by experts is often the only practical way to get useful corpora for CL. (Artstein and Poesio 2008: 590)

Domain specificity is also an issue when trying to evaluate a particular annotation scheme. Craggs and McGee Wood note that in discourse and dialogue studies, applying one scheme across several domains has been encouraged as a way of gauging the

reliability of the scheme. But as they point out, this evaluates the *process* (of interest in Content Analysis, not in CL) rather than the annotation scheme itself. The correct approach is to apply the scheme only within a single domain. In the situation with multiple domains "[a]ny differences in the results between corpora are a function of the variance between samples and not of the reliability of the coding scheme." (2005: 290) This is clearly one of the problems encountered in the results of the current study: the extreme range of inter-annotator agreement among corpus articles (and domains) has introduced the articles themselves as an unexpected factor in the analysis of the inter-annotator variation. Although as presented in detail in Sections 3.3.3.1-12 there are factors other than domain knowledge, such as variation in writing style, that play a part in the inter-article variation, the fact that none of the current annotators had expert-level understanding of any of the corpora domains led to uncertainty and variation. Even with multiple re-reads it is difficult for a non-expert to evaluate the authors' rhetoric in such highly technical content. Thus the task of comparing the performance of the two Models of argument has been made more difficult by needing to consider the variation by article, as well as by individual annotators (Tables 27, 30, 51-56).

It therefore seems clear that in order to eliminate the inter-annotator variation caused by the uncertainty and confusion on the part of non-expert annotators, experts in the biomedical domains represented in a corpus are required. The ideal would be to match experts to a particular sub-domain e.g., a specialist in Molecular Biology would annotate only articles from the *BMC Molecular Biology* journal. It might be worthwhile to have such experts annotate the articles from our final corpus and compare the results to those

reported here; this would provide some idea of how much of our current variation had been caused by our lack of biomedical expertise. Annotators with a major investment in a project based on their own research interests, including those who collaborate over the longer term, are certainly different from those hired briefly who are poorly paid and "typically have no stake in the end result" (Zaenen: 579). For the purposes of generating data that can be useful in Machine Learning, domain experts and those heavily invested in the study's outcome are the most likely to reduce both the errors and the systematic inter-annotator variation found in the results of the current study.

5.4 Annotation Process

This study, being on a small scale with only three annotators and over a single academic term, relied largely on inter-personal communication (emails) and collaborative discussion during the training period. The only 'official' written instructions were those found, along with the revised Models, in Appendices A, C, D and F. For larger scale projects with more annotators a detailed set of instruction guidelines becomes necessary. Although a list of examples can be helpful, it does not take the place of allowing annotators to train on 'real' data which are representative of the corpus being used. Although the RST project noted in Section 1.2 had an enormous amount of documentation (Carlson et al.) a more typical CL project is that of Wilbur et al. (2006) where they had six pages of instruction guidelines and eighteen pages of Appendices containing examples.

During training, feedback and discussion between annotators and project director(s) are crucial. Guidelines may need to be revised, and depending on the state of the Model(s) of argument being applied, changes to the coding scheme may be appropriate. In the ideal situation described above in 5.3, where there would be two sets of annotators – biomedical experts and experts in argument – they should at some point be trained together in order to be familiar with the different aspects of the annotation process. Individual annotators from each group could work together in pairs, thus sharing their respective expertise during the annotation process. If time allows, the annotators could record their thought processes involved when deciding on difficult cases and/or, as mentioned above, record the category they Trumped as well as the category they selected.

As long as the focus of annotation is on argument, I recommend the process used here: annotators should read the entire article, but annotate only the Discussion section. But again I stress, as in 5.2 above, that consultation with biomedical researchers is critical; we are assuming that analyzing the rhetoric will be useful to them (or at least some of them), but this should be confirmed with a variety of senior and junior researchers. One of the benefits of using biomedical expert annotators would be that they, unlike the three annotators in this study, would be able to evaluate the Model(s) being applied with some understanding of the end-user's requirements; their feedback could thus prove extremely useful in considering future revisions to the Models and the instruction guidelines.

5.5 Hedges

The results of this study indicate that lexical hedges could be used as possible cues for particular argument categories. Although the final corpus is relatively small, the striking commonality of the hedge distributions shown in Table 57 suggests that either the ten or fifteen most commonly occurring hedges are representative of those found across large electronic biomedical research corpora, and could be used as a preliminary 'gold-standard' list in the future for automated sentence/clause classification . As discussed in Section 4.3.3, under Model 1 the presence of a lexical hedge is a strong indicator of the ANALYSIS (5) category, especially if there is no in-text citation. In the presence of a citation and a verb in the past tense, a hedge most likely implies the CONTEXT (1) category. Under Model 2, 67.0% of hedges were found in either categories CLAIM (1) or QUALIFIER (4). Even though it is clear that both of these categories need to be redefined (see Section 4.2.1.2), it is worth examining where hedges occur in future studies using a revised Model 2, especially in relation to CLAIMS, the core of Toulmin's argument structure.

Lexical hedges are worth considering as category cues in any future studies of argument given that they are readily found by automated search tools. IE for biomedical texts is an active research area, and hedges are being used as indicators of speculation or lack of certainty. A recently reported project used a machine learning system to find the scope of lexical hedge cues in a corpus of medical and biological texts (Morante and Daelemans 2009). Another recent study in CL has used weakly supervised machine learning to classify hedged sentences in a corpus of papers in Genomics as either 'speculative' or

'non-speculative'; they use non-lexical hedges, a considerably more difficult task than using lexical items. It is interesting to note that some of their hedge types e.g., 'statement of speculative hypothesis', 'statement of knowledge paucity', resemble categories of rhetoric (Medlock and Briscoe 2007: 993-994). In future work allowing argument categorization at the clause rather than the sentence level will allow for finer-grained analysis; hedge distributions in such a study should be compared to the results of this project to see if categorizations remain similar.

5.6 Argument Type

Given that even after much collaboration among annotators and the development of a new set of Argument Types we did not all agree on Type for one of the twelve final corpus articles, I believe this approach of categorizing at the Discussion level should not be pursued. It seems that rather than being a useful 'macro' guide to selecting micro-level argument categories (under either Model of argumentation), it was another source of inter-annotator variation, and frustration at trying to identify the one 'correct' Type. My own experiences as an annotator and the results of this study suggest that some type of 'Trumping' system operating at the unit of annotation, such as those developed for Models 1 and 2, is more useful both as a guide for annotators and a tool to reduce inter-annotator agreement. It may be worth exploring in the future whether Argument Types could be used at the paragraph level, mid-way between the macro and micro units, as the paragraph generally comprises a particular aspect of the authors' argumentation. The utility of this approach would depend on the Argument Types identified and the Model of argument being applied.

5.7 Summary

This thesis has reported on an annotation project where two Models of argument/rhetoric were applied to the Discussion sections of a corpus of biomedical research articles. Model 1 is information-based and contains five categories (such as CURRENT RESULTS and ANALYSIS) and Model 2 is based on Toulmin's argument structure (1958/2003) and contains six categories (such as CLAIM and QUALIFIER). Each sentence was categorized under both Models by three annotators, myself (the project director) and two fourth-year Medical Sciences students at UWO. For a training period of several months, five 'practice' articles from the on-line *BMC*-series of journals were annotated; during that time collaboration and discussions among the three annotators led to revisions to both preliminary Models of argument. The revised Models 1 and 2 were then applied to the final corpus of twelve *BMC* articles; here annotators worked completely independently and submitted their data electronically. I then collected and organized all annotated data: twelve articles totalling 400 sentences for each of the two Models, for each of the three annotators.

Being one of the three annotators was crucial in my understanding of both the Models of argument and the corpus data; through applying the Models I could see where one Model was a better fit than the other for particular corpus data, and where specific rhetorical categories were not clearly defined and/or did not seem appropriate for our biomedical corpus. By comparing our annotations for each article sentence by sentence I was able to observe what kinds of corpus data were problematic for each of the Models, and to

provide detailed examples of the different types of inter-annotator variation e.g., legitimate subjective differences of opinion, misunderstandings of a Model's categories, errors, etc. Also, the fact that the training period had allowed me to become familiar with my two annotators meant that I had some insight into why their annotations differed from my own, and respect for the instances where their choices made it evident that they were right and I was wrong. Having performed such micro-analyses of our results gave me a clearer understanding when I stepped back to look at the bigger picture, of the many dimensions in the inter-annotator variation that were identified. In this thesis I have provided detailed breakdowns of this variation – by annotator, by Model, by Model categories, by inter-annotator agreement categories, and by corpus article. All variation found in this project – between annotators, Models, categories, corpus articles – can be of use not only in improving the Models applied here (see 5.1 above), but in increasing our understanding of what makes the annotation process harder or easier, and who is the best annotator to match with particular corpus data.

These analyses make it clear that under each Model some categories are major sources of inter-annotator disagreements, whereas others seem easier to agree on, and that revisions are required for both Models of argument before future applications. Although Model 1 'outperformed' Model 2 in terms of overall inter-annotator agreement, it is premature to rule out a Toulmin-based Model of argumentation. This is the first time Toulmin's model of argument has been applied to the rhetorical analysis of biomedical research texts, so it is not surprising that further modifications to Model 2 would be necessary. It may be that an adapted CLAIMS-based Model would be useful for the needs of biomedical researchers.

The most surprising finding – the enormous range of three-way agreement among articles – has shown the absolute necessity of using biomedical domain experts in any further annotation projects. Instead of two articles in e.g., Biochemistry, there needs to be a corpus of at least ten articles in the same domain, annotated, possibly in tandem with those skilled in rhetoric (see above), by experts in that field. Once the changes recommended in this thesis have been made, a decision, based in part on the current results, will need to be made on the most appropriate coefficient of agreement to employ in future studies; as discussed above, it is important to select a coefficient that is able to differentiate between different types of disagreements, rather than masking them. In addition, a reasonable threshold of agreement will need to be identified for each particular usage being made of the annotated data.

And lastly, the possibility that some of the current annotated corpus data might be of use in future work should not be ruled out. It may be that the data with three-way inter-annotator agreement, approximately 60% in the case of Model 1 and 40% under Model 2, could be used in future applications.

WORKS CONSULTED

Andrews, James, Rachel Richesson and Jeffrey Krischer. 2007. Variation of SNOMED CT Coding of Clinical Research Concepts among Coding Experts. *Journal of the American Medical Informatics Association*, Vol. 14(4), 497-506.

Artstein, Ron and Massimo Poesio. 2008. Inter-coder Agreement for Computational Linguistics. *Computational Linguistics*, Vol. 34(4), 555-596.

Bazerman, C. 1988. *Shaping written knowledge: the genre and activity of the experimental article in science*. Madison: University of Wisconsin Press.

Bellert, Irena & Paul Weingartner. 1982. On different characteristics of scientific texts as compared with everyday language texts. In *Sublanguage: studies of language in restricted semantic domains*, Richard Kittredge & John Lehrberger eds., 219-230. Berlin; New York: W. de Gruyter.

Cann, Ronnie. 1993. *Formal semantics, An introduction*. Cambridge University Press.

Carletta, Jean. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, Vol. 22(2), 249-254.

Carlson, L., D. Marcu, & M. E. Okurowski. 2001. Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, Denmark.

Craggs, Richard and Mary McGee Wood. 2005. Evaluating Discourse and Dialogue Coding Schemes. *Computational Linguistics*, Vol. 31(3), 289-296.

Creswell, C., K. Forbes, E. M., R. Prasad, A. Joshi, & B. Webber. 2002. The Discourse Anaphoric Properties of Connectives. In *Proceedings of DAARC2002*, Edicões Colibri.

Crompton, Peter. 1997. Hedging in Academic Writing: Some Theoretical Problems. In *English for Specific Purposes* Vol. 16(4), 271-287.

Dalianis, H. and P. Johannesson. 1997. Explaining Conceptual Models using Toulmin's argumentation and RST. In *Conceptual Modeling – ER '97*, 215-228. Berlin/Heidelberg: Springer.

DiEugenio, Barbara and Michael Glass. 2004. The Kappa Statistic: A Second Look. *Computational Linguistics*, Vol. 30(1), 95-101.

DiMarco, C., F. W. Kroon, & R. E. Mercer. 2005. Using hedges to classify citations in scientific articles. In *Computing Attitude and Affect in Text: Theory and*

Applications, J. G. Shanahan, Y. Qu, J. Wiebe eds., 247—264. Dordrecht, The Netherlands: Springer.

Docherty, Michael and Richard Smith. 1999. The case for structuring the discussion of scientific papers. In *British Medical Journal* 318, 1224-1225.

Foss, Sonja K., Karen A. Foss and Robert Trapp. 2002. *Contemporary Perspectives on Rhetoric*. Prospect Heights, Illinois: Waveland Press.

Fox, John and Sanjay Modgil. 2006. From arguments to decisions: extending the Toulmin view. In *Arguing on the Toulmin model: New essays in argument analysis and evaluation*. David Hitchcock and Bart Verheij eds., 273-288. Dordrecht: Springer.

Garssen, Bart. 2001. Argument Schemes. In *Crucial concepts in argumentation theory*, F. van Eemeren ed., 81-100. Amsterdam: Amsterdam University Press.

Graves, Heather. 2005. *Rhetoric in(to) Science : Style as Invention in Inquiry*. Cresskill, New Jersey: Hampton Press.

Graves, Heather and Roger Graves. 2007. *A Strategic Guide to Technical Communication*. Peterborough: Broadview Press.

Gross, Alan. 1990. *The Rhetoric of Science*. Cambridge, Mass/London, England: Harvard University Press.

Gross, Alan & Ray Dearin. 2003. *Chaim Perelman*. Albany: State University of New York Press.

Hahn, Udo and Joachim Wermter. 2006. Levels of Natural Language Processing for Text Mining. In *Text Mining for Biology and Biomedicine*, Sophia Ananiadou and John McNaught eds. Boston/London: Artech House.

Halliday, M.A.K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Harris, Randy Allen. 1997. *Landmark Essays on the Rhetoric of Science: Case Studies*. R Harris ed. Mahwah, N.J.: Hermagoras Press.

He, X. and C. DiMarco. 2005. Using Lexical Chaining to rank protein-protein interaction in biomedical text. Poster presentation, BioLINK, Detroit, June 2005.

Hearst, Marti. 1999. Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. 3-10. College Park, MD.

Hearst, Marti. 2003. *What is Text Mining?* www.jaist.ac.jp/~bao/MOT-Ishikawa/FurtherReadingNo1.pdf (Accessed August 3, 2009)

Henkemans, A. Francisca. 2001. Argument Structures. In *Crucial concepts in argumentation theory*, F. van Eemeren ed., 101-134. Amsterdam: Amsterdam University Press.

Hiz, Henry. 1982. Specialized languages of biology, medicine and science and connections between them. In *Sublanguage: studies of language in restricted semantic domains*, Richard Kittredge & John Lehrberger eds., 206-212. Berlin; New York: W. de Gruyter.

Horton, Richard. 1995. The rhetoric of research. In *British Medical Journal* 310, 985-987.

Hunter, Lawrence and K. Bretonnel Cohen. 2006. Biomedical Language Processing: Perspective What's Beyond PubMed? *Mol Cell*, 21(5), 589-594.

Hyland, Ken. 1996. Talking to the Academy: Forms of Hedging in Science Research Articles. In *Written Communication* Vol. 13(2), 251-281.

Hyland, Ken. 1998. *Hedging in Scientific Research Articles*. Amsterdam/Philadelphia: John Benjamins.

Jenicek, Milos. 2006. How to read, understand, and write 'Discussion' sections in medical articles: An exercise in critical thinking. *Med Sci Monit*, 12(6), SR28-SR36.

Kim, Jin-Dong, Tomoko Ohta and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics* 9: 10.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to its methodology*. Beverly Hills/London: Sage.

Krippendorff, Klaus. 2004. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research* Vol. 30(3), 411-433.

Kumar, C. Anil and V. Vishnu. 2001. PubMed Central: A phenomenal advance in electronic publishing. *Current Science* Vol. 81(1).

Langer, Hagen, Harald Lüngen and Petra Saskia Bayerl. 2004. *Text Type structure and logical document structure*. ACL Workshop on Discourse Analysis, 2004.

Locke, David. 1992. *Science as Writing*. New Haven: Yale University Press.

Lu, Zhiyong, Michael Bada, Philip V. Ogren, Bretonnel Cohen and Lawrence Hunter. 2006. Improving Biomedical Corpus Annotation Guidelines. *The Joint BioLINK and 9th Bio-Ontologies Meeting 2006.*

Mann, William and Sandra Thompson. 1987. Rhetorical structure theory: Description and construction of text structures. In *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*, G. Kempen ed., 85-95. Dordrecht: Martinus Nijhoff.

Marcu, Daniel. 2000. *The theory and practice of discourse parsing and summarization.* Cambridge: MIT Press.

Medlock, Ben and Ted Briscoe. 2007. Weakly Supervised Learning for Hedge Classification in Scientific Literature. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*; 992-999.

Mizuta, Yoko and Nigel Collier. 2004. An Annotation Scheme for a Rhetorical Analysis of Biology Articles. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004).*

Mizuta, Yoko, Tony Mullen and Nigel Collier. 2004. Zone Identification in Biology Articles as a Basis for Information Extraction. In *Proceedings of the International Joint Workshop on NLP in Biomedicine and its Applications (JNLPBA).*

Mizuta, Yoko, Anna Korhonen, Tony Mullen and Nigel Collier. 2005. Zone Analysis in Biology Articles as a Basis for Information Extraction. *International Journal of Medical Informatics* Vol. 75(6), 468-487. Elsevier.

Mooney, Raymond and Razvan Bunescu. 2005. Mining Knowledge from Text Using Information Extraction. *SIGKDD Explorations*, Vol. 7(1), 3-10.

Morante, Roser and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the Workshop on BioNLP*; 28-36.

Morris, Jane and Graeme Hirst. 1991. Lexical cohesion, the thesaurus, and the structure of text. *Computational linguistics* 17(1), 21-48.

Myers, Greg. 1990. *Writing Biology: Texts in the Social Construction of Scientific Knowledge*. Madison: University of Wisconsin Press.

Myers, Greg. 1994. Narratives of science and nature in popularizing molecular genetics. In *Advances in written text analysis*, Malcolm Coulthard ed., 179-190. London; New York: Routledge.

Perelman, Chaim. 1971. *Logique et argumentation.* Bruxelles: Presses universitaires de Bruxelles.

Perelman, Chaim and Lucie Olbrechts-Tyteca. 1969. *The new rhetoric: A treatise on argumentation*. John Wilkinson and Purcell Weaver, trans. University of Notre Dame Press.

Prelli, Lawrence J. 1989. *A Rhetoric of Science: Inventing Scientific Discourse*. University of South Carolina Press.

Reed, Chris and Glenn Rowe. 2006. Translating Toulmin diagrams: theory neutrality in argument representation. In *Arguing on the Toulmin model: New essays in argument analysis and evaluation*. David Hitchcock and Bart Verheij eds., 341-358. Dordrecht: Springer.

Shankar, R.D., S. W. Tu and M. A. Musen. 2006. Medical Arguments in an Automated Health Care System. In *AAAI spring symposium technical report*, Stanford Medical School.

Shatkay, Hagit, Fengxia Pan, Andrey Rzhetsky and W. John Wilbur. 2008. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, Vol. 24 (18), 2086-2093.

Sintonen, Matti. 2004. Argument, Inference and Reasoning – Integrating Induction and Deduction. In *Induction and Deduction in the Sciences*, Friedrich Stadler ed., 121-134. Dordrecht/Boston/London: Kluwer.

Siegel, Sidney and N.J. Castellan Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*, 2nd edition. McGraw-Hill.

Skelton, John. 1988. The care and maintenance of hedges. In *ELT Journal [English Language Teaching]* Vol. 42(1), 37-43.

Skelton, John R. and Sarah J.L. Edwards. 2000. The function of the discussion section in academic medical writing. In *British Medical Journal* 320, 1269-1270.

Stoddard, Sally. 1991. *Text and texture: Patterns of cohesion*. Norwood, N.J.: Ablex.

Swales, John. 1990. *Genre analysis: English in academic and research settings*. Cambridge: CUP.

Teufel, Simone. 1999. *Argumentative Zoning : Information Extraction from Scientific Text*. PhD thesis, School of Cognitive Science, University of Edinburgh.

Teufel, Simone and Mark Moens. 1999. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In *Advances in Automatic*

Text Summarization, Inderjeet Mani and Mark Maybury, eds., 155-171. Cambridge, MA: MIT Press.

Teufel, Simone, Jean Carletta and Mark Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Eighth Meeting of the European Chapter of the Association for Computational Linguistics (EACL-99)*, 110-117.

Teufel, Simone and Mark Moens. 2000. What's yours and what's mine : Determining intellectual attribution in scientific text. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural language Processing and Very Large Corpora.*

Teufel, Simone and Mark Moens. 2002. Summarizing Scientific Articles : Experiments with Relevance and Rhetorical Status. *Computational Linguistics* 28(4), 409-445.

Tindale, Christopher W. 1999. *Acts of Arguing: A Rhetorical Model of Argument.* Albany: State University of New York Press.

Toulmin, Stephen E. 1953. *The Philosophy of Science: An Introduction.* London/Toronto: Hutchinson's University Library.

Toulmin, Stephen E. 1958/2003. *The Uses of Argument.* Cambridge University Press.

van Eeemeren, Frans. 2001. The State of the Art in Argumentation Theory. In *Crucial concepts in argumentation theory*, F. van Eemeren ed., 11-26. Amsterdam: Amsterdam University Press.

van Eeemeren, Frans and Rob Grootendorst. 2004. *A Systematic Theory of Argumentation, The pragma-dialectal approach.* Cambridge: CUP.

Watters, S., B. McInnes, D. McKoskey, T Miller, D. Boley, M. Gini, W. Schuler, A. Polukeyeva, J. Gundel, S. Pakhomov, G. Savova. 2005. Using Volunteers to Annotate Biomedical Corpora for Anaphora Resolution. *American Association for Artificial Intelligence Spring Symposium.*

White, Barbara E. 2005a. *L'analyse des zones et les zones argumantales : une comparaison de deux cadres d'analyse rhétoriques pour les textes scientifiques.* M.S., University of Western Ontario.

White, Barbara E. 2005b. *The Ideal Unit of Annotation for the Rhetorical Analysis of Scientific Texts.* Bilingual Workshop in Theoretical Linguistics, UWO, December 9, 2005.

White, Barbara E. 2006. *Une approche pratique au développement d'un modèle amélioré de l'analyse rhétorique pour l'annotation des textes tirés des corpus biomédicaux.* M.S., University of Western Ontario.

Wiebe, Janyce M., Rebecca F. Bruce and Thomas P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *ACL99: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD; 246-253.

Wilbur, W. John, Andrey Rzhetsky and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and construction. *BMC Bioinformatics* 2006, 7:356.

Zaenen, Annie. 2006. Mark-up: Barking up the wrong tree. *Computational Linguistics* Vol. 32(4), 577-580.

Zeiger, Mimi. 2000. *Essentials of Writing Biomedical Research Papers, 2nd Edition.* McGraw- Hill, Health Professions Division.

APPENDIX A: Instructions to Annotators - January 2008

Each reader has their own unique response to any text; there are no 'correct' or 'incorrect' annotations. Although we are trying to analyze the writers' presentation of their argumentation, we cannot be inside their heads; ultimately each reader's view of the rhetoric is somewhat subjective. I am interested in finding out how different people understand/perceive the rhetoric of the same material: Where do they agree/disagree? Where do they have difficulties evaluating the choice of category? I also want input regarding the application of different models: Do they seem appropriate for these (*BioMed*) data? Are some categories more problematic or unclear? Should the text of the Discussion be exhaustively annotated?

READING THE ARTICLE
Did you read the whole document?
Did you read some or all of it more than once?
If only parts, which ones? E.g., Abstract and Discussion
In what order did you read the sections?
Was the abstract necessary and/or useful?
Did you review the charts, tables, graphics etc.?
    How important were they to your understanding of the document?
    How important were they to the authors' rhetoric?
Did you skim/read the Bibliography?
Did you follow any of the in-text citations to the Bibliography for information/clarification?
Track your time for your reading (taking into account the above questions) – roughly, not to the second!
How much time did it take you to annotate the Discussion section?

CONTENT

How familiar/comfortable are you generally with the field of the article (e.g., Microbiology)?
How familiar are you with the material in the particular article?
Did you find that the most significant/persuasive rhetoric was in the Discussion?

If you could not categorize a particular sentence, was this because:
    You felt more than 1 category could apply?
        None of the categories applies?
    You were not clear on the meaning of the text:
        Not familiar with the material/terminology?
        Not sure what the authors' point is?
        Not clear how it relates to the surrounding sentences?
            To the core argumentation?
                i.e., the argument or discourse seems to lack coherence

What was the most difficult part of this task? Why?

APPENDIX B: Articles in Training Corpus

T1) Comparative 3-D Modeling of tmRNA.
Research Article
BMC Molecular Biology
www.biomedcentral.com/1471-2199/6/14

T2) Carvacrol and p-cymene inactivate Escherichia coli O157:H7 in apple juice.
Research Article
BMC Microbiology
www.biomedcentral.com/1471-2180/5/36

T3) Redesigned and chemically-modified hammerhead ribozymes with improved activity and serum stability
Research Article
BMC Chemical Biology
www.biomedcentral.com/1472-6769/4/1

T4) Association study of genetic variants of pro-inflammatory chemokine and cytokine genes in systemic lupus erythematosus
Research Article
BMC Medical Genetics
www.biomedcentral.com/1471-2350/7/48

T5) Localization of plasma membrane t-SNAREs syntaxin 2 and 3 in intracellular compartments
Research Article
BMC Cell Biology
www.biomedcentral.com/1471-2121/6/26

APPENDIX C: Revised Model 1

1) CONTEXT

- Background to the current experiment; generally accepted knowledge in the field
- Work currently being carried out by other researchers
- Descriptions of, or discussions related to, earlier research projects
- Statements of results from any work previous to the current study
- Existing or previous debate in the field; open research questions/issues
- Motivation for the current experiment

2) METHOD

- Descriptions of methods, processes, tools etc. used in current study
- Basis for choice of above
- Descriptions of experimental design
- *Only* material specific to the current study

3) CURRENT RESULTS

- Statements of what they found in their current study
- May include references to tables, graphics, etc.

4) RESULTS COMPARED

- Their current results are similar to, consistent with previous results
- Their current results contrast/are inconsistent with previous results

5) ANALYSIS

- Suggest why something did/did not happen
- Possible interpretations/implications of current or previous results
- Speculation of any sort by current authors
- Indicate significance of their present findings
- Limitations of their experiment/results
- Implications for the field and suggestions for future work

CATEGORY 'TRUMPING' IN THE FACE OF COMPLEXITY/UNCERTAINTY

(3) Trumps (4):	Especially if the new results in (4) are not stated elsewhere (Focus is on new results rather than old)
(2) through (5) Trump (1):	Focus on information gain rather than history
(5) Trumps (3),(4):	If the statement is critical to their argumentation

APPENDIX D: Revised Model 2

0) EXTRANEOUS

- Background to the current experiment; undisputed facts in the field
- Any of the following that are not directly related to a CLAIM:
  - Current debate in the field
  - Motivation for the current experiment
  - Assumptions made going into the experiment
  - Statements related to the methodology

1) CLAIM

- Proposition put forward based on analysis and interpretation of results
- Not simply a result or an observation; what the results *mean*

Major CLAIM: always based on current results

Minor CLAIM: may be based on previous work (external evidence)

2) GROUNDS

- Data: internal evidence drawn from the authors' current study
- Material used to support a CLAIM

3) WARRANT/BACKING

- Understanding of the problem based on external evidence
- Specific data and information from other (external) studies
- Additional support for a CLAIM

4) QUALIFIER

- Possible explanations for their data
- Alternate explanations for diverging results
- Compare and contrast with external evidence
  - May act as a bridge from external evidence to a current CLAIM

5) PROBLEM IN CONTEXT

- Implications of their *current* study for the future of their field
- How the current results shed light on or alter the path of future research
  - Ways the CLAIM qualifies or impacts the larger problems
  - New directions for additional research on the broader issues

CATEGORY 'TRUMPING' IN THE FACE OF COMPLEXITY/UNCERTAINTY

(1) Trumps (2) through (5): The central focus of this model is to identify CLAIMS
(2) Trumps (3): Internal evidence is at the core of the current Argument
(4) Trumps (3): (4) makes external evidence relevant to their CLAIM

APPENDIX E: Articles in Final Corpus

C1) Expression and localization of estrogenic type 12 17ß-hydroxysteroid dehydrogenase in the cynomolgus monkey
Research Article
BMC Biochemistry
www.biomedcentral.com/1471-2091/8/2

C2) Limited proteolysis of human histone deacetylase I
Research Article
BMC Biochemistry
www.biomedcentral.com/1471-2091/7/22

C3) Mapping of A_1 conferring resistance to the aphid *Amphorophora idaei* and *dw* (dwarfing habit) in red raspberry (*Rubus idaeus* L.) using AFLP and microsatellite markers
Research Article
BMC Plant Biology
www.biomedcentral.com/1471-2229/7/15

C4) An informatics search for the low-molecular weight chromium-binding peptide
Research article
BMC Chemical Biology
www.biomedcentral.com/1472-6769/4/2

C5) Discovery of chemically induced mutations in rice by TILLING
Methodology Article
BMC Plant Biology
www.biomedcentral.com/1471-2229/7/19

C6) Protein expression of G-protein inwardly rectifying potassium channels (GIRK) in breast cancer cells.
Research Article
BMC Physiology
www.biomedcentral.com/1472-6793/6/8

C7) An immune response in the bumblebee, *Bombus terrestris* leads to increased food consumption
Research Article
BMC Physiology
www.biomedcentral.com/1472-6793/6/6

C8) Differential development of neuronal physiological responsiveness in two human neural stem cell lines
Research Article
BMC Neuroscience
www.biomedcentral.com/1471-2202/8/36

C9) Degradation of the LDL receptors by PCSK9 is not mediated by a secreted protein acted upon by PCSK9 extracellularly
Research Article
BMC Cell Biology
www.biomedcentral.com/1471-2121/8/9

C10) The effect of ABCA1 gene polymorphisms on ischaemic stroke risk and relationship with lipid profile
Research article
BMC Medical Genetics
www.biomedcentral.com/1471-2350/8/30

C11) Serum procalcitonin elevation in critically ill patients at the onset of bacteremia caused by either gram negative or gram positive bacteria
Research article
BMC Infectious Diseases
www.biomedcentral.com/1471-2334/8/38

C12) Reference genes for normalization of gene expression studies in human osteoarthritic articular cartilage
Research article
BMC Molecular Biology
www.biomedcentral.com/1471-2199/9/17

APPENDIX F: Instructions to Annotators - Final Corpus, April 2008

1) You will annotate all twelve articles in the Final Corpus list. Read an article, abstract first, then the remainder of the article (skip the Methods section if you feel it is not necessary for your understanding of the argumentation).
2) Select one of the four Argument Types, whichever seems to best suit their main line of argumentation in the Discussion section. Remember that we are categorizing arguments, not types of experiments. Enter the code for the Argument Type (1-4) in the appropriate spot at the top of the WORD version of the Discussion section.
3) Also in this section enter the Model number (1 or 2) under which you are annotating.
4) Annotate the Discussion section under the most recent versions of Model 1 and Model 2 by highlighting each sentence with the appropriate colour, or no colour if Category (0) under Model 2. Leave at least 1 white space between sentences, and do not annotate the text which indicates an in-text citation e.g., '[35, 37]'. In situations where you are uncertain: the sentence is grammatically or rhetorically complex; it seems to fit in 2 different categories at the same time; it does not seem to fit any category; it is difficult to clearly understand the content of the sentence, you must choose an annotation category. Use the 'Trumping' guidelines (included with the Models) to assist you with particular conflicting situations, but otherwise use your understanding of argumentation as we have been discussing it, and especially of the key concepts behind our 2 models. Make use of the 'References' section at the end of the article as necessary to help clarify the rhetoric of statements alluding to external evidence. Note that under Model 2, a 'Claim' may be based on previous evidence, but it must be a current statement i.e., it must be given in the present tense (a past 'Claim' may possibly be categorized as 'Warrant/Backing').
5) Create one SPSS sheet for each of the twelve articles. Enter the data regarding the article, paragraph, and sentence numbers in the appropriate columns of the SPSS file. Enter the code for the category of each sentence under both Models; in situations where you have difficulty deciding between 2 categories (for whatever reason, and under either Model) enter your 'best' choice in the category column.
6) For each sentence check for the occurrence of any form of the lexical items in the 'HEDGES' list of March 30, 2008. Enter data only if you found a hedge in that sentence i.e., 'missing' (empty) cells are OK here. Note that for verb forms you should record the verb tense form as found in the text e.g., *appears, seemed, speculating* etc. For nouns, enter the form as found in the text i.e., either singular or plural. Enter these (1 per cell) in the order in which they occur in the sentence, up to a maximum of 3.
7) Save the SPSS sheet with a file name beginning with your initials, and which includes the article number (as on the Corpus list, 1-12) as well as some word(s) referencing the content of the paper.
8) Save the colour-annotated Discussion sections with file names as in (7) above, but with the addition of the Model number under which it is annotated.

APPENDIX G

The effect of *ABCA1* gene polymorphisms on ischaemic stroke risk and relationship with lipid profile

BMC Medical Genetics 2007, **8:**30doi:10.1186/1471-2350-8-30

The electronic version of this article is the complete one and can be found online at: http://www.biomedcentral.com/1471-2350/8/30

C10 Annotator: BW MODEL #: 1 ARGUMENT TYPE: 4

Discussion

The *ABCA1* gene is known to have a crucial role in lipid metabolism [29,30]. Mutations or polymorphisms in this gene are known to cause dyslipidemia such as low HDL-C and thus predispose to atherosclerosis [31,32]. Several polymorphisms of the *ABCA1* gene have been investigated for their association with CAD [33-35]. Other works have reported that CAD patients who are carriers of R219K allele had less severe atherosclerosis [31] and overall lower risk of CAD [36]. Hindrikovics *et al* recently reported a higher frequency of R219K in controls than in Hungarian stroke patients and suggested a protective role for this polymorphism [25]. By contrast, our control group had a higher "R219" allele frequency, while the stroke population had a higher "219K" allele frequency. The "219K" allele frequency is similar to that reported in other studies of Irish and other Scottish populations [36]. Two SNPs, P1648L and T1555I, were not polymorphic in our population.

While the R219K -G1051A- (A or "K" variant) has been associated with decreased TG, increased HDL and subsequently a lower risk for atherosclerotic progression, in contrast the R allele has been associated with vascular disease [31]. This has not been confirmed in our study, although in our stroke population those with R219K "22" genotype (AA) had a higher level of LDL and the "K allele" carriers had a lower TG. Lee *et al* also reported lower TG in the carriers of 219K variant and this finding was replicated in our population ($p = 0.006$). The HDL level showed no significant difference among different R219K genotypes.

It is also of interest that a protective role for the 219K allele has not been confirmed by all studies. Ethnic background or other environmental factors may weaken the link with HDL-C levels. However, in three other European populations in contrast to a Japanese one, R219 has been constantly the wild type allele [37].

Haplotype analysis can provide additional power in association studies in complex disease [38]. Results using different programs are usually consistent, but sometimes there are minor variations [6,39]. We performed haplotype analysis in the remaining four SNPs, and only the 2211 and 1211 haplotypes were more frequent in cases ($p = 0.05$). Only a small proportion of individuals carried these haplotypes, thus the result should be interpreted with caution.

We found an association between LDL levels and *ABCA1* genotype, but not with HDL. Epidemiological studies of *ABCA1* polymorphisms and HDL levels suggest that only 10% of HDL level variation maybe explained by this gene [40] and thus our study may not have been large enough to detect this. Other studies have shown an association between the R219K polymorphism and MI, but no association between haplotype arrangements and MI. Polymorphisms in the promoter region (C-564T) and in the coding region (R1587K) have shown an association with ApoA-I levels but these have not been associated with vascular disease. Another study has suggested that rare alleles with major phenotypic effects contribute significantly to low plasma HDL [41]. Although lipid measurements were made early after admission, possible confounders include the acute lipid changes that occur after acute stroke. The lipid levels reported in our study are similar to those values on the morning after admission reported by Dyker, Weir and Lees in patients after acute stroke [42]. The changes in lipids post stroke remain controversial, but further studies of changes in lipid profile will be difficult because of the early introduction of statin therapy on the basis of studies such as SPARCL [43].

APPENDIX H

The effect of *ABCA1* gene polymorphisms on ischaemic stroke risk and relationship with lipid profile

BMC Medical Genetics 2007, **8**:30doi:10.1186/1471-2350-8-30

The electronic version of this article is the complete one and can be found online at: http://www.biomedcentral.com/1471-2350/8/30

C10 Annotator: BW MODEL #: 2 ARGUMENT TYPE: 4

Discussion

The *ABCA1* gene is known to have a crucial role in lipid metabolism [29,30]. Mutations or polymorphisms in this gene are known to cause dyslipidemia such as low HDL-C and thus predispose to atherosclerosis [31,32]. Several polymorphisms of the *ABCA1* gene have been investigated for their association with CAD [33-35]. Other works have reported that CAD patients who are carriers of R219K allele had less severe atherosclerosis [31] and overall lower risk of CAD [36]. Kindrikovics et al recently reported a higher frequency of R219K in controls than in Hungarian stroke patients and suggested a protective role for this polymorphism [25]. By contrast, our control group had a higher "R219" allele frequency, while the stroke population had a higher "219K" allele frequency. The "219K" allele frequency is similar to that reported in other studies of Irish and other Scottish populations [36]. Two SNPs, P1648L and T1555I, were not polymorphic in our population.

While the R219K -G1051A- (A or "K" variant) has been associated with decreased TG, increased HDL and subsequently a lower risk for atherosclerotic progression, in contrast the R allele has been associated with vascular disease [31]. This has not been confirmed in our study, although in our stroke population those with R219K "22" genotype (AA) had a higher level of LDL and the "K allele" carriers had a lower TG. Clee *et al* also reported lower TG in the carriers of 219K variant and this finding was replicated in our population (p = 0.006). The HDL level showed no significant difference among different R219K genotypes.

It is also of interest that a protective role for the 219K allele has not been confirmed by all studies. Ethnic background or other environmental factors may weaken the link with HDL-C levels. However, in three other European populations in contrast to a Japanese one, R219 has been constantly the wild type allele [37].

Haplotype analysis can provide additional power in association studies in complex diseases [38]. Results using different programs are usually consistent, but sometimes there are minor variations [6,39]. We performed haplotype analysis in the remaining four SNPs, and only the 2211 and 1211 haplotypes were more frequent in cases [p = 0.05]. Only a small proportion of individuals carried these haplotypes, thus the result should be interpreted with caution.

We found an association between LDL levels and *ABCA1* genotype, but not with HDL. Epidemiological studies of *ABCA1* polymorphisms and HDL levels suggest that only 10% of HDL level variation maybe explained by this gene [40] and thus our study may not have been large enough to detect this. Other studies have shown an association between the R219K polymorphism and MI, but no association between haplotype arrangements and MI. Polymorphisms in the promoter region (C-564T) and in the coding region (R1587K) have shown an association with ApoA-I levels but these have not been associated with vascular disease. Another study has suggested that rare alleles with major phenotypic effects contribute significantly to low plasma HDL [41]. Although lipid measurements were made early after admission, possible confounders include the acute lipid changes that occur after acute stroke. The lipid levels reported in our study are similar to those values on the morning after admission reported by Dyker, Weir and Lees in patients after acute stroke [42]. The changes in lipids post stroke remain controversial, but further studies of changes in lipid profile will be difficult because of the early introduction of statin therapy on the basis of studies such as SPARCL [43].

APPENDIX I

The effect of *ABCA1* gene polymorphisms on ischaemic stroke risk and relationship with lipid profile

BMC Medical Genetics 2007, **8**:30doi:10.1186/1471-2350-8-30

The electronic version of this article is the complete one and can be found online at: http://www.biomedcentral.com/1471-2350/8/30

C10 Annotator: JH MODEL #: 1 ARGUMENT TYPE: 3

Discussion

The *ABCA1* gene is known to have a crucial role in lipid metabolism [29,30]. Mutations or polymorphisms in this gene are known to cause dyslipidemia such as low HDL-C and thus predispose to atherosclerosis [31,32]. Several polymorphisms of the *ABCA1* gene have been investigated for their association with CAD [33-35]. Other works have reported that CAD patients who are carriers of R219K allele had less severe atherosclerosis [31] and overall lower risk of CAD [36]. Kindrikovics *et al* recently reported a higher frequency of R219K in controls than in Hungarian stroke patients and suggested a protective role for this polymorphism [25]. By contrast, our control group had a higher "R219" allele frequency, while the stroke population had a higher "219K" allele frequency. The "219K" allele frequency is similar to that reported in other studies of Irish and other Scottish populations [36]. Two SNPs, P1648L and T1555I, were not polymorphic in our population.

While the R219K -G1051A- (A or "K" variant) has been associated with decreased TG, increased HDL and subsequently a lower risk for atherosclerotic progression, in contrast the R allele has been associated with vascular disease [31]. This has not been confirmed in our study, although in our stroke population those with R219K "22" genotype (AA) had a higher level of LDL and the "K allele" carriers had a lower TG. Clee *et al* also reported lower TG in the carriers of 219K variant and this finding was replicated in our population ($p = 0.00$). The HDL level showed no significant difference among different R219K genotypes.

It is also of interest that a protective role for the 219K allele has not been confirmed by all studies. Ethnic background or other environmental factors may weaken the link with HDL-C levels. However, in three other European populations in contrast to a Japanese one, R219 has been constantly the wild type allele [37].

Haplotype analysis can provide additional power in association studies in complex diseases [38]. Results using different programs are usually consistent, but sometimes there are minor variations [6,39]. We performed haplotype analysis in the remaining four SNPs, and only the 2211 and 1211 haplotypes were more frequent in cases ($p = 0.05$). Only a small proportion of individuals carried these haplotypes, thus the result should be interpreted with caution.

We found an association between LDL levels and *ABCA1* genotype, but not with HDL. Epidemiological studies of *ABCA1* polymorphisms and HDL levels suggest that only 10% of HDL level variation maybe explained by this gene [40] and thus our study may not have been large enough to detect this. Other studies have shown an association between the R219K polymorphism and MI, but no association between haplotype arrangements and MI. Polymorphisms in the promoter region (C-564T) and in the coding region (R1587K) have shown an association with ApoA-I levels but these have not been associated with vascular disease. Another study has suggested that rare alleles with major phenotypic effects contribute significantly to low plasma HDL [41]. Although lipid measurements were made early after admission, possible confounders include the acute lipid changes that occur after acute stroke. The lipid levels reported in our study are similar to those values on the morning after admission reported by Baker, Weir and Lees in patients after acute stroke [42]. The changes in lipids post stroke remain controversial, but further studies of changes in lipid profile will be difficult because of the early introduction of statin therapy on the basis of studies such as SPARCL [43].

APPENDIX J

The effect of *ABCA1* gene polymorphisms on ischaemic stroke risk and relationship with lipid profile

BMC Medical Genetics 2007, **8**:30doi:10.1186/1471-2350-8-30

The electronic version of this article is the complete one and can be found online at: http://www.biomedcentral.com/1471-2350/8/30

C10 Annotator: JH MODEL #: 2 ARGUMENT TYPE: 3

Discussion

The *ABCA1* gene is known to have a crucial role in lipid metabolism [29,30]. Mutations or polymorphisms in this gene are known to cause dyslipidemia such as low HDL-C and thus predispose to atherosclerosis [31,32]. Several polymorphisms of the *ABCA1* gene have been investigated for their association with CAD [33-35]. Other works have reported that CAD patients who are carriers of R219K allele had less severe atherosclerosis [31] and overall lower risk of CAD [36]. Andrikovics *et al* recently reported a higher frequency of R219K in controls than in Hungarian stroke patients and suggested a protective role for this polymorphism [25]. By contrast, our control group had a higher "R219" allele frequency, while the stroke population had a higher "219K" allele frequency. The "219K" allele frequency is similar to that reported in other studies of Irish and other Scottish populations [36]. Two SNPs, P1648L and T1555I, were not polymorphic in our population.

While the R219K -G1051A- (A or "K" variant) has been associated with decreased TG, increased HDL and subsequently a lower risk for atherosclerotic progression, in contrast the R allele has been associated with vascular disease [31]. This has not been confirmed in our study, although in our stroke population those with R219K "22" genotype (AA) had a higher level of LDL and the "K allele" carriers had a lower TG. Cline *et al* also reported lower TG in the carriers of 219K variant and this finding was replicated in our population (p = 0.006). The HDL level showed no significant difference among different R219K genotypes.

It is also of interest that a protective role for the 219K allele has not been confirmed by all studies. Ethnic background or other environmental factors may weaken the link with HDL-C levels. However, in three other European populations in contrast to a Japanese one, R219 has been constantly the wild type allele [37].

Haplotype analysis can provide additional power in association studies in complex diseases [38]. Results using different programs are usually consistent, but sometimes there are minor variations [6,39]. We performed haplotype analysis in the remaining four SNPs, and only the 2211 and 1211 haplotypes were more frequent in cases (p = 0.05). Only a small proportion of individuals carried these haplotypes, thus the result should be interpreted with caution.

We found an association between LDL levels and *ABCA1* genotype, but not with HDL. Epidemiological studies of *ABCA1* polymorphisms and HDL levels suggest that only 10% of HDL level variation maybe explained by this gene [40] and thus our study may not have been large enough to detect this. Other studies have shown an association between the R219K polymorphism and MI, but no association between haplotype arrangements and MI. Polymorphisms in the promoter region (C-564T) and in the coding region (R1587K) have shown an association with ApoA-I levels but these have not been associated with vascular disease. Another study has suggested that rare alleles with major phenotypic effects contribute significantly to low plasma HDL [41]. Although lipid measurements were made early after admission, possible confounders include the acute lipid changes that occur after acute stroke. The lipid levels reported in our study are similar to those values on the morning after admission reported by Dyker, Weir and Lees in patients after acute stroke [42]. The changes in lipids post stroke remain controversial, but further studies of changes in lipid profile will be difficult because of the early introduction of statin therapy on the basis of studies such as SPARCL [43].

APPENDIX K

The effect of *ABCA1* gene polymorphisms on ischaemic stroke risk and relationship with lipid profile

BMC Medical Genetics 2007, **8:**30doi:10.1186/1471-2350-8-30

The electronic version of this article is the complete one and can be found online at: http://www.biomedcentral.com/1471-2350/8/30

C10 Annotator: KP MODEL #: 1 ARGUMENT TYPE: 3

Discussion

The *ABCA1* gene is known to have a crucial role in lipid metabol [29,30]. ations or polymorphisms in this gene are known to cause dyslipidemia such as low HDL-C and thus predispose to atherosclerosis [31,32]. Several polymorphisms of the *ABCA1* gene have been investigated for their association with CAD [33-35]. Other works have reported that CAD patients who are carriers of R219K allele had less severe atherosclerosis [31] and overall lower risk of CAD [36]. Andrikovics *et al* recently reported a higher frequency of R219K in controls than in Hungarian stroke patients and suggested a protective role for this polymorphism [25]. By contrast, our control group had a higher "R219" allele frequency, while the stroke population had a higher "219K" allele frequenc. The "219K" allele frequency is similar to that reported in other studies of Irish and other Scottish population [36]. Two SNPs, P1648L and T1555I, were not polymorphic in our population.

While the R219K -G1051A- (A or "K" variant) has been associated with decreased TG, increased HDL and subsequently a lower risk for atherosclerotic progression, in contrast the R allele has been associated with vascular disease [31]. This has not been confirmed in our study, although in our stroke population those with R219K "22" genotype (AA) had a higher level of LDL and the "K allele" carriers had a lower T. Lee *et al* also reported lower TG in the carriers of 219K variant and this finding was replicated in our population (p = 0.006). The HDL level showed no significant difference among different R219K genotypes.

It is also of interest that a protective role for the 219K allele has not been confirmed by all studies. Ethnic background or other environmental factors may weaken the link with HDL-C levels. However, in three other European populations in contrast to a Japanese one, R219 has been constantly the wild type allele [37].

Haplotype analysis can provide additional power in association studies in complex diseases [38]. Results using different programs are usually consistent, but sometimes there are minor variations [6,39]. We performed haplotype analysis in the remaining four SNPs, and only the 2211 and 1211 haplotypes were more frequent in cases (p = 0.05). Only a small proportion of individuals carried these haplotypes, the the result should be interpreted with caution.

We found an association between LDL levels and *ABCA1* genotype, but not with HDL. Epidemiological studies of *ABCA1* polymorphisms and HDL levels suggest that only 10% of HDL level variation maybe explained by this gene [40] and thus our study may not have been large enough to detect this. Other studies have shown an association between the R219K polymorphism and MI, but no association between haplotype arrangements and MI. Polymorphisms in the promoter region (C-564T) and in the coding region (R1587K) have shown an association with ApoA-I levels but these have not been associated with vascular disease. Another study has suggested that rare alleles with major phenotypic effects contribute significantly to low plasma HDL [41]. Although lipid measurements were made early after admission, possible confounders include the acute lipid changes that occur after acute stroke. The lipid levels reported in our study are similar to those values on the morning after admission reported by Dyker, Weir and Lees in patients after acute stroke [42]. The changes in lipids post stroke remain controversial, but further studies of changes in lipid profile will be difficult because of the early introduction of statin therapy on the basis of studies such as SPARCL [43].

APPENDIX L

The effect of *ABCA1* gene polymorphisms on ischaemic stroke risk and relationship with lipid profile

BMC Medical Genetics 2007, **8**:30doi:10.1186/1471-2350-8-30

The electronic version of this article is the complete one and can be found online at: http://www.biomedcentral.com/1471-2350/8/30

C10 Annotator: KP MODEL #: 2 ARGUMENT TYPE: 3

Discussion

The *ABCA1* gene is known to have a crucial role in lipid metabolism [29,30]. Mutations or polymorphisms in this gene are known to cause dyslipidemia such as low HDL-C and thus predispose to atherosclerosis [31,32]. Several polymorphisms of the *ABCA1* gene have been investigated for their association with CAD [33-35]. Other works have reported that CAD patients who are carriers of R219K allele had less severe atherosclerosis [31] and overall lower risk of CAD [36]. Andrikovics *et al* recently reported a higher frequency of R219K in controls than in Hungarian stroke patients and suggested a protective role for this polymorphism [25]. In contrast, our control group had a higher "R219" allele frequency, while the stroke population had a higher "219K" allele frequency. The "219K" allele frequency is similar to that reported in other studies of Irish and other Scottish populations [36]. Two SNPs, P1648L and T1555I, were not polymorphic in our population.

While the R219K -G1051A- (A or "K" variant) has been associated with decreased TG, increased HDL and subsequently a lower risk for atherosclerotic progression, in contrast the R allele has been associated with vascular disease [31]. This has not been confirmed in our study, although in our stroke population those with R219K "22" genotype (AA) had a higher level of LDL and the "K allele" carriers had a lower TG. Klee *et al* also reported lower TG in the carriers of 219K variant and this finding was replicated in our population (p = 0.006). The HDL level showed no significant difference among different R219K genotypes.

It is also of interest that a protective role for the 219K allele has not been confirmed by all studies. Ethnic background or other environmental factors may weaken the link with HDL-C levels. However, in three other European populations in contrast to a Japanese one, R219 has been constantly the wild type allele [37].

Haplotype analysis can provide additional power in association studies in complex diseases [38]. Results using different programs are usually consistent, but sometimes there are minor variations [6,39]. We performed haplotype analysis in the remaining four SNPs, and only the 2211 and 1211 haplotypes were more frequent in cases (p = 0.05). Only a small proportion of individuals carried these haplotypes, thus the result should be interpreted with caution.

We found an association between LDL levels and *ABCA1* genotype, but not with HDL. Epidemiological studies of *ABCA1* polymorphisms and HDL levels suggest that only 10% of HDL level variation maybe explained by this gene [40] and thus our study may not have been large enough to detect this. Other studies have shown an association between the R219K polymorphism and MI, but no association between haplotype arrangements and MI. Polymorphisms in the promoter region (C-564T) and in the coding region (R1587K) have shown an association with ApoA-I levels but these have not been associated with vascular disease. Another study has suggested that rare alleles with major phenotypic effects contribute significantly to low plasma HDL [41]. Although lipid measurements were made early after admission, possible confounders include the acute lipid changes that occur after acute stroke. The lipid levels reported in our study are similar to those values on the morning after admission reported by Dyker, Weir and Lees in patients after acute stroke [42]. The changes in lipids post stroke remain controversial, but further studies of changes in lipid profile will be difficult because of the early introduction of statin therapy on the basis of studies such as SPARCL [43].