Rhetorical Patterns in Citations across Disciplines and Levels of Participation

Andreas Karatsolis

Massachusetts Institute of Technology | USA

Abstract: Writing researchers have long attempted to classify and describe patterns of citation and source use both to describe disciplinary differences, and to identify discourse-level characteristics of new knowledge production. The analysis of large corpora has provided great insights about the formal characteristics of citations, but little information about their rhetorical nature, which we know from interview studies as central to the understanding of source use practices. This study reports on an attempt to understand and describe patterns of source use across disciplines, genres and levels of participation through systematic verbal data analysis of documents produced by sixteen participants in expert/novice pairs (faculty advisor/doctoral advisee) from four disciplines (Computer Science, Chemical Engineering, Materials Science Engineering and Humanities and Social Sciences). The results of this analysis showed that, despite some disciplinary differences, all participants used similar patterns of reference use, namely elaboration, evaluation and relation to one's own work.

Keywords: citation studies, novice/expert, verbal data analysis, disciplinarity

1. Introduction

Since the end of the 20th century and certainly in the beginning of the 21st, academic work has seen some dramatic changes, driven both by external pressures (increased enrollments, accountability demands and relations with industry), as well as changes in the demands for research and the knowledge academics produce (Coaldrake and Stedman, 1999). In fact, Gibbons described a shift in the mode of knowledge production from individualized and local knowledge systems, to team-based, collaborative distributed systems drawing knowledge together from diverse sources (1998). Such a shift was made possible, to some extent, by information technologies that allowed researchers unprecedented access to vast knowledge repositories, as well as the most up-to-date information, not only from their field, but across disciplines. This access was the result of new forms of publishing (e-journals and online conference proceedings) as well as new approaches and tools to abstracting and indexing services. Presently, the "acquisition and storage" component of new scientific knowledge (see the King, MacDonald and Roderer model from 1981) is being executed almost exclusively through online databases, at least for the libraries of major universities and research centers in the US and Europe.

Given this newfound access to document repositories, a fascinating "dataset" for writing researchers was, naturally, the published journals and proceedings from different disciplines, which include not only the published text, but also descriptive metadata (author, date, keywords etc.), and, of course, citations; the latter, despite their flat presentation structure, have a great potential to trace idea relationships in fields, knowledge networks and disciplinary patterns of attribution and acknowledgment. Given the importance of citation analysis first as a tool for journal evaluation (see Garfield, 1972) and, consequently, the academic reappointment and promotion process, researchers from multiple disciplines have attempted to understand the nature of citation practices. Several studies have used author co-citation analysis to map the intellectual base of disciplines such as Macroeconomics (McCain, 1984), Organizational Behavior (Mc Culnan et al., 1990), Information and Library Science (Ding et al., 1999), and even scholarly communication in Sociology of Science (Karki, 1996). A common theme running underneath most of these studies is that citation practices and citation functions are often discipline-specific and sometimes highly individualized, as researchers often make source selection and reference decisions based on content and context considerations. In fact, as early as the mid-1980s Cronin had argued that citation is a social act and any co-occurrence or textual analysis cannot fully capture its inherently rhetorical nature (1984). Cronin proposed that citation is not a unit, but an event, which is very difficult to lay bare as one would have to step into the author's head to understand the functional, social or political motivations behind a single instance of citation.

Around that time, the rhetorical nature of texts was being explored in new ways, including rhetorical moves in introduction sections of journal articles (Swales and Najar, 1987) or author positioning moves (such as disagreement) in academic discourse (Hunston, 1993). One line of research that developed as an extension of these efforts is related to the identification of specific discourse patterns in different genres as they emerge from corpus analyses, with citation a particular pattern of interest. The work of Ken Hyland (1999, 2001, 2002) stands out, since he has analyzed citation patterns through several measures, including reporting verbs, integral/non-integral citations and self-mention. His conclusions were important for the understanding of disciplinary practices, especially in terms of understanding the knowledge construction process of different groups and the epistemological and social conventions of disciplines. Using a similar method to analyze citations from corpora, Thomson and Tribble (2001) examined the difference between integral and non-integral citations in doctoral theses from agricultural economics and agricultural botany. They found differences not only between these seemingly similar disciplines, but also between levels of participation, where novice writers seemed to use a limited range of citation types. More recent work in writing studies using citation corpora have even compared citation use across languages (Soler-Monreal & Gil-Salom, 2011) or across document types such as research articles and master's theses (Samraj, 2013). However, such studies do not provide insight into the reasons or the motivation behind differences in use of citation types. On the other hand, more ethnographic approaches in the form of one or two case studies of graduate students writing (Connor and Kramer, 1998 or Nielsen and Rocco, 1999), which do focus on the motivations behind citation use, reveal only a very narrow portion of the phenomenon.

Within this larger context, several Writing Studies researchers embarked on projects which would uncover some of the motivations behind the citation choices academic authors make by asking the authors themselves, instead of attempting to reconstruct meaning by studying the text for markers or patterns of use. Most notable are studies by Wang and Soergel (1998) and Harwood (2005) who attempted to identify dimensions and categories of source use decisions by interviewing authors from Agricultural Sciences, and Computer Science and Sociology respectively. In the first study, the motivation was to explore the reasons participants gave in interviews regarding the whole process of conducting research, including writing and publication. The reasons participants offered were clearly linked to citation "functions" (or what Swales would call "rhetorical moves"), though Wang and White originally termed them "contributions" (1997). Similarly, Harwood's study (2005) used qualitative interviews in which participants described their motivation behind each citation instance in order to arrive at a taxonomy of citation functions. Additionally, he confirmed the multifunctional nature of citations, as most of the citations his participants discussed involved multiple functions.

A similar effort was undertaken by the author of this article (Karatsolis, 2005), with a focus on examining expert/novice differences in citation use by interviewing advisor-

advisee pairs of participants from four disciplines (Materials Science Engineering, Chemical Engineering, Humanities and Social Sciences and Computer Science) using a discourse-based interview model (see Odell, 1985). This work was motivated by the recognition that, even within closely knit research groups where authorship was shared between graduate students and the principal investigator, the understanding of the functions of citations and the value of sources varied greatly and was a point of constant contention between experts and novices. The overall goal of the work, then, became the development of a visualization framework, called *Kairion*, which would allow experts and novices to position their work in relation to the literature and the common rhetorical patterns of use (see Karatsolis, 2011 for more details).

In order, however, for the discourse-based interviews to follow the protocol proposed by Odell (1985), in which participants not only are presented with an instance of discourse from their own writing (such as a citation) and are asked to comment on it, but they are asked to explain if the use of an alternative would have made a difference, a substantial corpus from the participants had to be collected and analyzed. These documents, which were solicited directly by the sixteen participants in the study as their most representative and recent scholarly work, constitute the common dataset upon which this special issue focused. This paper will report on the methodology for this type of coding and analysis, following Geisler's verbal data analysis framework (2003), from the published or unpublished documents collected by the participants. In addition, the results of this hand coding, which have not been presented in the past as they primarily served to inform the discourse-based interviews, will be discussed, with an eye towards disciplinary differences, as well as levels of participation in the field.

2. Methodology

As briefly mentioned in the previous section, participants were recruited from four different departments representing distinct disciplinary approaches (Materials Science Engineering, Chemical Engineering, Humanities and Social Sciences and Computer Science) in advisor/advisee pairs. The pairs were typically a dissertation advisor and an advanced doctoral student, at some stage of the process of completing their Ph.D. As is common the United States, the doctoral students were either in the Exam stage, in which they had completed their coursework and a written examination on three or four areas of disciplinary interest, the ABD stage, in which they had completed and defended a dissertation proposal, or the Ph.D. stage in which they had already written dissertation chapters and published journal articles, but had not defended their dissertation. Table 1 shows the distribution of participants for both the advisees and the advisors, who were distributed based on the traditional academic ranks of Assistant, Associate and Full Professor in the US.

The recruitment process involved a first meeting with the advisee, in which their verbal consent to participate was solicited. If their response was positive, a meeting

with their advisor was scheduled, in which their written consent to formally participate in the study was obtained (IRB approved by Rensselaer Polytechnic Institute as Protocol #519). At the same time, a request was made for documents they had authored, which they typically provided in hard copy or pdf form. Finally, in a follow-up meeting with the advisee, their consent was formally obtained as well as any documents on which they were the primary author, including unpublished dissertation chapters and prospectuses. After the recruitment, for almost one academic year, observations of advisor-advisee meetings, either as one-to-one meetings or as an advisor-research group, were scheduled, especially for meetings in which the documents I had obtained were reviewed. In total, about 25 advisors and advisees originally agreed to participate in the study. However, only 16 participants, in 8 pairs across the four disciplines, eventually completed all parts of the study. This paper will report on the results of the analysis of the documents from the 16 participants, as they were coded to help support the final discourse-based interviews with each participant.

Table 1: Disciplinary affiliation and position of study participants

	Exam	ABD	Ph.D.	Assistant	Associate	Full
MSE		☑	☑	☑	☑	
CS		☑	☑		☑	☑
ChemEng	☑	☑				☑☑
HSS	☑		☑	☑	☑	

Most of the documents were provided in print or pdf form, so they had to be turned into editable text using Optical Character Recognition software (OmniPro 12). This part of the process also involved some manual cleaning of the data, especially since character recognition was not possible for equations and figures, which often included references and had to be added to the textual corpus. In total, 31 documents were obtained: 24 published journal articles, 5 dissertation chapters from three different participants, and two candidacy proposals. Each complete document was then segmented into sentences, and then transferred into a formatted Excel spreadsheet which included coding categories and descriptors. The total corpus for all sixteen participants was 3,804 sentences, a relatively small number for corpora work, but rather large for descriptive discourse analysis standards, especially for serving the goal of producing discourse-based interview protocols. Each sentence occupied one cell in a spreadsheet within a column that included all the text to be coded from the document. In that sense, coders could read the document in a linear manner, but following a column from top to bottom instead of paragraphs or pages of text. The spreadsheets were de-identified and validation rules were put in place to ensure that only the possible codes for each coding pass (e.g. 0/1 or A/B/C) could be entered.

As previously explained, the original purpose of the coding of the documents was to design structured discourse-based interview protocols for each participant, based on the

patterns identified in the analysis. For this reason, the coding category definition stage was extensive and went through multiple iterations, as the coding categories had to be clear and easily identifiable not only to the coders, but also to the authors/participants themselves. In fact, the first iteration of the coding scheme was based on feedback from the second coders and the dissertation committee members, but the next iteration also included feedback from the pilot participants. Four participants (two from the advisor and two from the advisee category) were eventually excluded from the final study results, as their documents were used for the pilot coding and refinement of coding category descriptions.

In fact, coders were initially presented with the pilot data set and were asked to offer recommendations for refining the category descriptions and the examples. In the first training iteration, the coders had to identify instances of citation and use marginal annotations to explain if they found the citation fall within a specific pattern of use (for example integral vs non-integral citation, author mention or single vs multiple reference). A session with the principal investigator followed to discuss coding decisions and categories. The pilot participants were then also consulted on some of the more complicated decisions or in cases where there was disagreement between the coder and the principal investigator.

Based on the feedback from this initial process and private conversations with Lee Odell, a preliminary coding scheme was constructed. The decision was to focus on categories that did not necessarily have distinct markers, avoiding ones that, even though easier to code, wouldn't be generative for the interviews that would follow because they were based on disciplinary formalisms (e.g. integral vs non-integral citations). Two coders, one doctoral and one master's student, were trained by the principal investigator, first by being presented with the coding scheme and examples, and then by coding a sample corpus from documents provided by pilot participants (excluded from the final analysis). The two coders then met with the principal investigator to adjudicate any splits in coding decisions and refine the coding scheme and examples.

The final coding scheme included four identifiable patterns (or moves) of relationship between the authored text and the cited sources:

1. **Reference**: *any instance where there is an explicit or implicit reference to a source, regardless of the presence or type of citation*. This was the easiest category to code as most citation styles require a very distinct and explicit reference to be made when a source is used; all sentences which included the name of an author, a parenthetical citation or a bracketed numerical reference automatically qualified. In cases where the reference to a source continued into a subsequent sentence (such as when the author provided details from a study or discussed its implications), the next sentence was also coded for reference. In addition, the more general pattern of single or multiple source reference was coded, as well as instances of integral

citation (where the author name was used in the sentence); however, the disciplinary differences in integral and non-integral citations were so pronounced that they were only used for the subsequent interviews, but the results will not be reported in this paper.

2. **Evaluation**: *any instance where there is an explicit or implicit evaluation of the cited source.* Only sentences which had been previously coded as "reference" could also be coded for this category. Coding this category required a decision as to whether the author was passing some judgment on the value of the source for the field or other sources or ideas/theories. If the author simply reported on information presented in the cited source(s) by using phrases such as "the author says/observes etc.," "this methodology was introduced," "increases have been reported," "these sources [3-6] use a similar methodology/framework," "experimental work has been conducted [2-6]," "it is reported in the literature [1-5]," the sentence was not coded as "evaluation."

3. **Elaboration**: *any instance where there is explicit or implicit elaboration on the information or ideas being presented from a cited source.* Even if a sentence did not include an explicit citation, it was examined for the possibility of being an extension or elaboration from a previous sentence which included a reference. In several cases, the citation for the source being referred to was either preceding or following the sentence where details or extended information about the source was presented. These sentences were also coded as evaluation. Similarly to the "Evaluation" category, if the author simply presented a single point or discrete piece of information from the cited source(s) by using neutral reporting phrases, we did not code for "elaboration."

4. **Relation to current project**: *any instance where there is explicit mention of the ideas or information presented in the cited source as they relate to the current project (the one that the author is describing/presenting).* Coding for this category sometimes required a closer attention to the key concepts presented in the research, but clear discourse markers of how the author built on or differentiated from other published sources were often present.

A more detailed explanation of the coding categories, along with the examples offered to coders can be found in the Appendix.

To measure reliability, a data set consisting of 5% of the total data, but almost 12% of the sentences which had been coded as Reference, was prepared. The dataset was drawn equally from all disciplines and all levels of participation. The two coders, who had already been trained with pilot data, received the dataset and were instructed to code for all three remaining categories, including the subcategories for each. For the category of Evaluation, agreement between coders in this category was .86 or .69 corrected using Cohen's Kappa. For the category of Elaboration, agreement between coders was .84 or .76 corrected using Cohen's Kappa. For the category of Relation to

Project, the second coder agreed with the first coder 96% of the time, but because of the small number of instances in the category, the adjusted reliability was 0.66 using Cohen's Kappa. For all three categories of analysis, the Cohen's Kappa was 0.785.

Given the large number of coding decisions, even with a relatively small corpus, and the added complication of some decisions being dependent on previous coding decisions, such reliability figures are very positive.

3. Results

As was mentioned in the introduction, the analysis of the documents provided by the 16 participants was originally aimed at identifying basic patterns of source use that the participants would clearly acknowledge, so as to formulate a series of questions for the discourse-based interviews that they would have no difficulty responding to. In that sense, the larger goal of this analysis was not to arrive at conclusive results about the differences in genres, disciplines, levels of participation, or intertextual connections, primarily because of its limited scope (participants and number of documents). However, as the analysis that follows will show, the four categories which we coded for can provide us with a complex model of citation patterns, which can possibly provide insights into differences between disciplines, across levels of participation or even the various ways of incorporating information that arises from other texts.

Given the built-in comparisons of the study in terms of levels of disciplinary participation (expert vs. novice source use patterns), disciplinary rhetorical moves involving source use (for the four disciplines analyzed), we also performed a ranked ANOVA test for discipline x participation. We also looked at the impact of genre (journal articles, dissertation chapters and thesis prospectuses). The remainder of this section has looks at these three areas in turn, with a larger discussion synthesizing the findings in the next section.

3.1 Analysis for patterns by levels of participation

For the purposes of this study, the obvious research question was if there are differences in source use patterns between advisors and advisees; however, results will also be reported based on the six participation levels presented in Table 1 (from full professors who had more than ten years in the field to doctoral students who had just defended their candidacy proposal, as is customary in US institutions for students who have completed the doctoral coursework and their comprehensive exams). None of the differences by level of participation reached the level of significance.

In terms of overall number of references for the complete corpus, advisees used more references than their advisors (see Figure 1), mostly more single source references (26% of total) than their advisors (16.8% of total). The multiple source references were about the same frequency (6.9% and 6.4% respectively). Since the corpus includes 2003 sentences by advisees versus 1793 sentences by advisors, a simple calculation

shows that, despite the fewer references overall, advisors used a slightly higher ratio of multiple sources (0.281) compared to advisees (0.264).
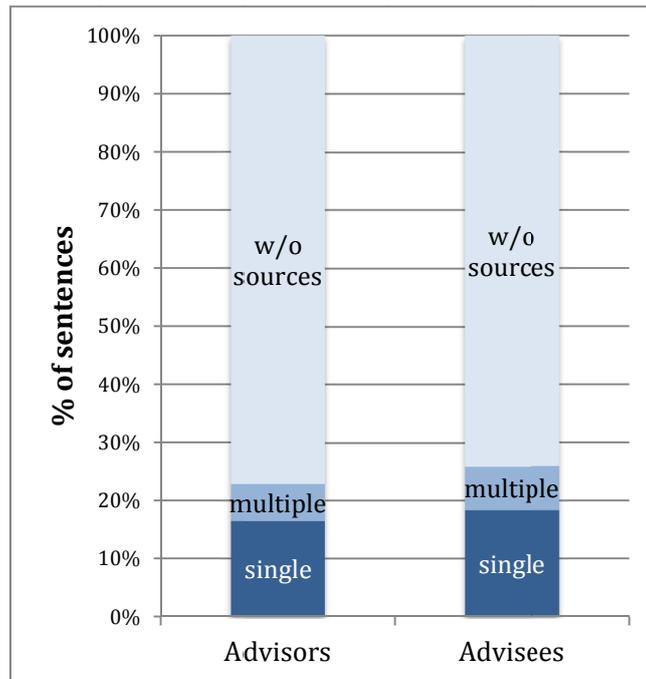
Figure 1: Overall Reference across disciplines.

If we examine in greater detail the sixteen participants in advisor/advisee pairs, first in terms of single source references, we see that for most pairs advisees have a higher percentage of single source use out of the total number of sentences in their documents, except for pair #6 from Materials Science Engineering (see Figure 2, where in this and all subsequent figures advisees are denoted by a red square, whereas advisors by a blue diamond shape).

The picture is quite different when examining multiple source use (Figure 3). In the two advisor/advisee pairs from the Humanities and Social Sciences (H&SS), advisees are referencing multiple sources within a single sentence more often, whereas in Materials Science and Engineering (MSE), the reverse seems to be the case. For the other two disciplines, Chemical Engineering (Chem. Eng.) and Computer Science (Comp. Sci.), there is no clear pattern between the pairs, though we have to say that the instances of multiple source use are so few in our corpus that a much larger study might provide us with more meaningful results. Overall, it seems that advisees use considerably more references in their documents, although the use of multiple sources varies between advisor/advisee pairs.
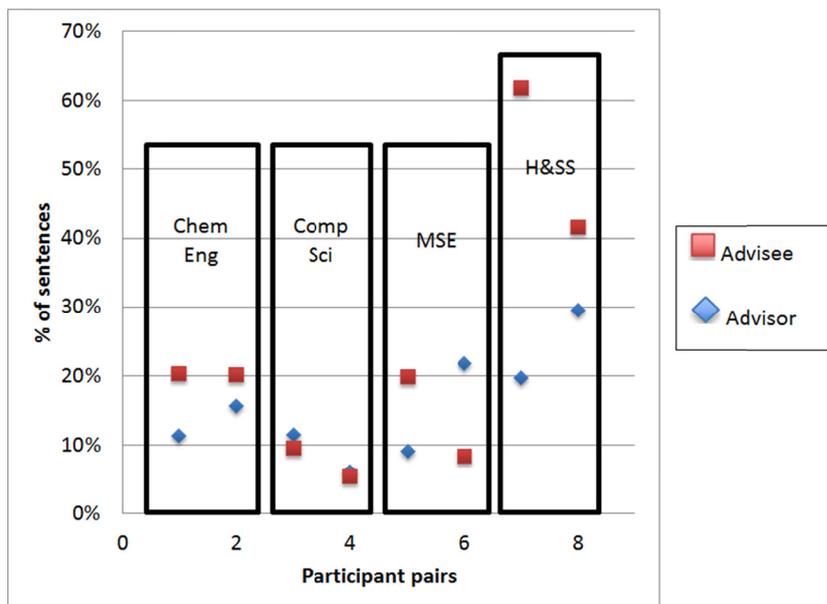
Figure 2: Comparison of single source reference by pairs.
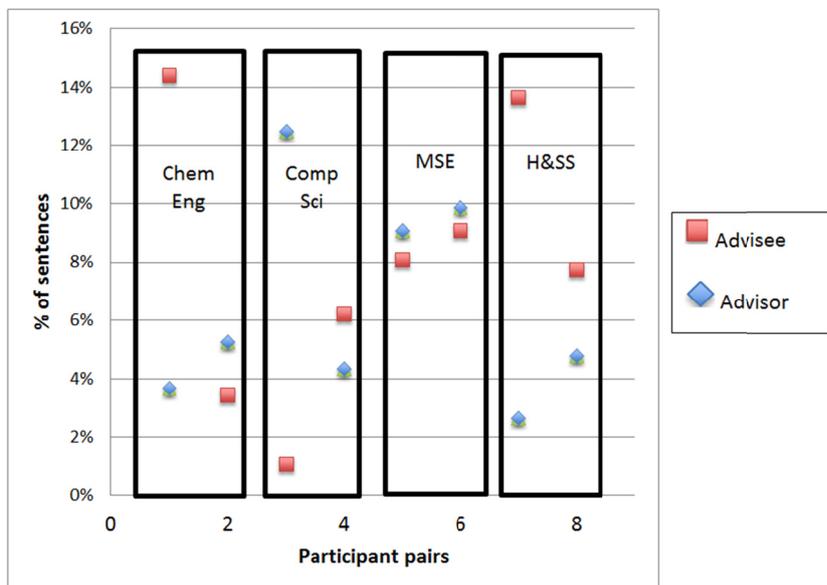
Figure 3: Comparison of multiple source references by pairs.

More interesting patterns begin to emerge as we analyze the other coding categories of Elaboration, Evaluation and Relation to Project. Figure 4 presents the overall results for the category of Elaboration. This total represents the total number of sentences that were coded for this dimension, which, for the most part, does not include sentences that had no citation as determined in the first coding. Only on occasion, a coder did code a sentence without citation for this dimension. This time we present the percentage of sentences coded as Elaboration in relation to the total sentences originally coded as Reference or adjacent to a sentence coded as Reference.
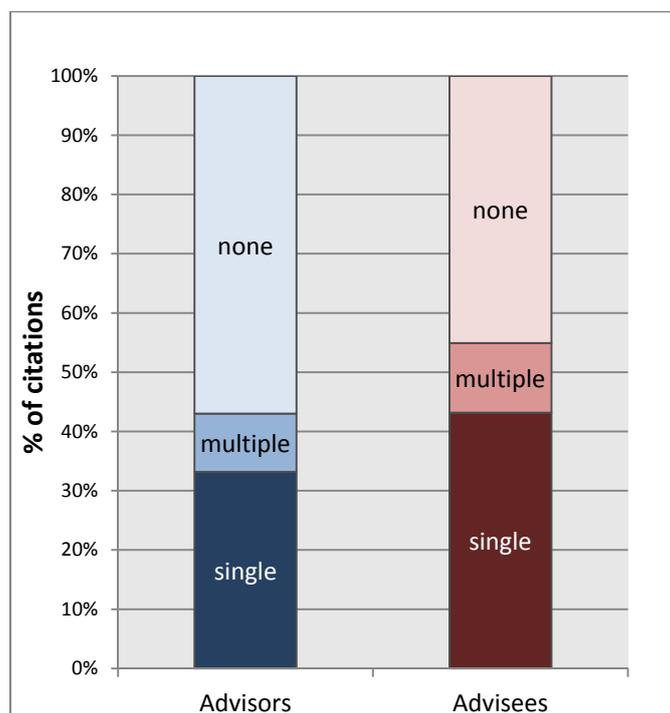
Figure 4: Overall Elaboration Patterns.

Again, advisees used proportionately more Elaboration of the referenced sources than their advisors (42% vs 54%), a statement true for 6 of the 8 pairs (see Figure 5). In addition, advisees use elaboration more for both in single source citations (43% vs 33% for advisors; true for 6 of 8 pairs) and multiple source citations (9% vs 6%; true for 7 of 8 pairs).
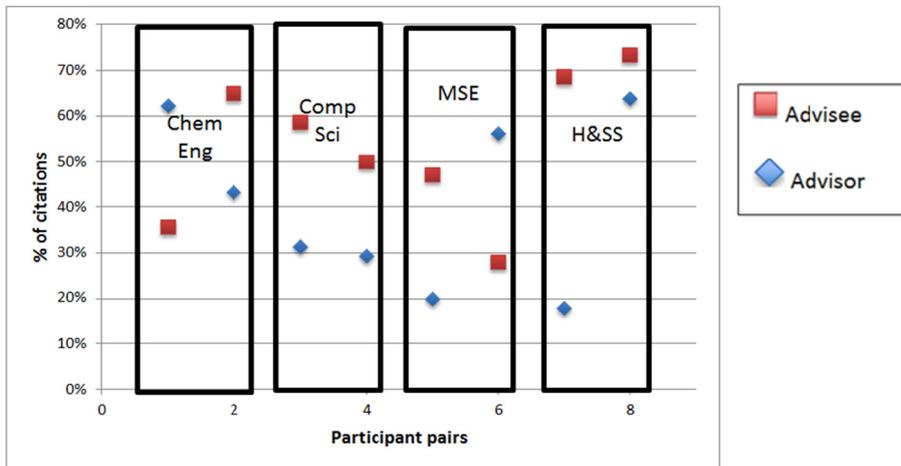
Figure 5: Comparison of Elaboration by participant pairs.

For the third category of analysis, Evaluation, the results (see Figure 6) followed a similar pattern as in the previous two categories: advisees do proportionately more evaluation of the citations than do their advisors (16% vs 9%; true for 7 of 8 pairs). In this category, only sentences previously coded as Reference could also be coded as Evaluation. Additionally, more Evaluation is done by advisees for both single source citations (10% vs 7%; true of 7 of 8 pairs) and multiple source citations (6% vs 2%; true of 7 of 8 pairs). Especially for multiple source evaluation by advisors, the percentage was so low (2% of all references) because several of them did not have any instances of evaluation of multiple sources within a single sentence.
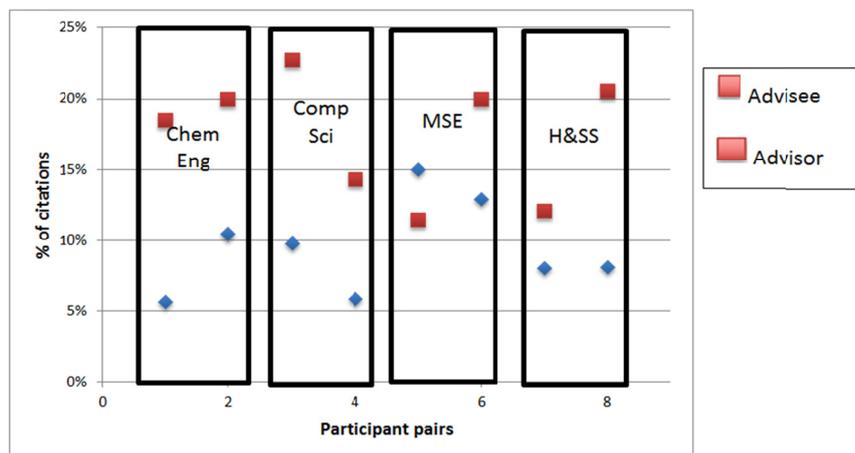
Figure 6: Comparison of Evaluation by participant pairs.

Finally, the analysis of the coding based on the fourth category, Relation to Project, yielded a quite different outcome than the previous three: Advisees place somewhat *fewer* citations in relation to their work as their advisors (11% vs 15%; true of 4 of 8 pairs). In terms of single or multiple source use, advisors place proportionately more of their single source citations in relation to their work than their advisees (12% vs 7%; true of 7 of 8 pairs). However, advisees place a few more of their multiple source citations in relation to their work than do their advisors (7% vs 9%; true for 5 of 8 pairs). Figure 7 shows the percentages for advisors and advisees, while Figure 8 shows the distribution of single source references in Relation to Project by pair.
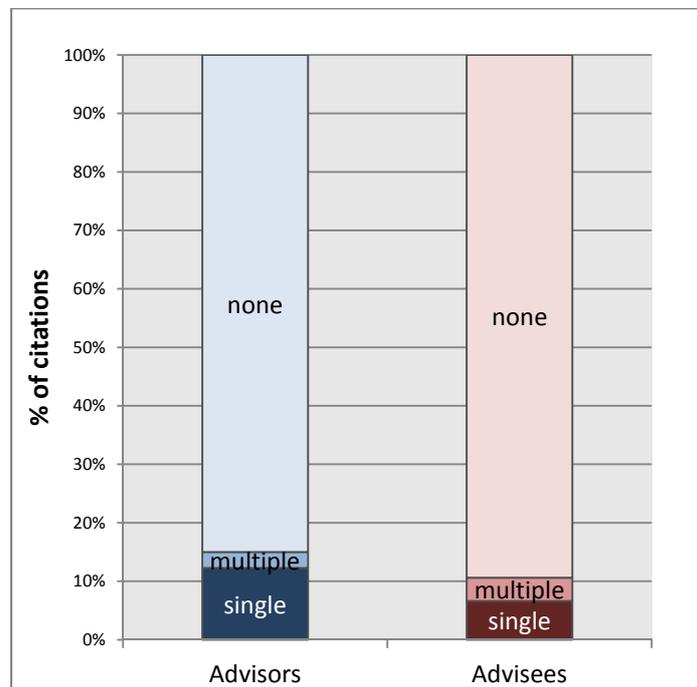
Figure 7: Overall Relation to Project Patterns.

As Figure 8 shows, all advisors used more instances of single source references in which they explicitly connected the source to their own work or project. Only the advisors for pairs 2 and 7 did not use this pattern at all, but their advisees did not either. In truth, coding for this pattern can be difficult because the connections might not necessarily be explicit or might be at a concept level, but there is a clear trend for advisors to exhibit this pattern more than advisees.
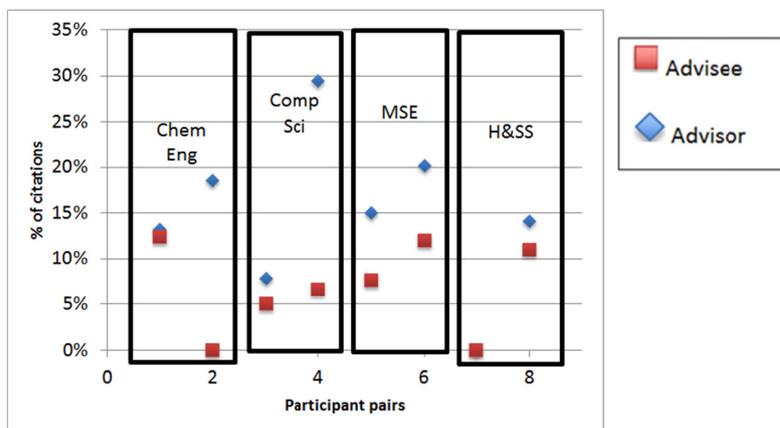
Figure 8: Single Source Citations coded for Relation to Project.

To explore this finding further, we analyzed the results of the third coding layer, in which we tried to identify if the reference was aimed at showing how the project at hand was building on or differing from previous work. The results showed most of the difference between advisors and advisees coming from the sub-category "building on the source." In fact, advisors had almost two times more instances of explicit statements showing how their own work or project builds on a source than advisees did (Figure 9). On the other hand, advisees and advisors had almost the same number of instances of explicitly showing how their works differs from a source (1.1% vs 0.9% respectively).
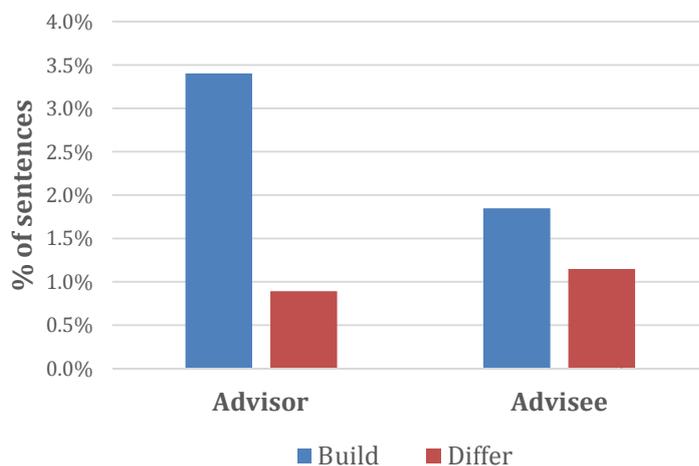
Figure 9: Comparison of Relation to Project patterns.

Continuing the analysis in terms of the full range of disciplinary participation, we expected to see some differences in the analytical categories, especially between students in the candidacy stage and full professors – the two ends of the spectrum in terms of years of participation in the discipline. Figure 10 presents the results from the three analytical categories (Evaluation, Elaboration and Relation to Project by participation status in the discipline, from graduate students having just completed their dissertation prospectus defense to full professors (from left to right).
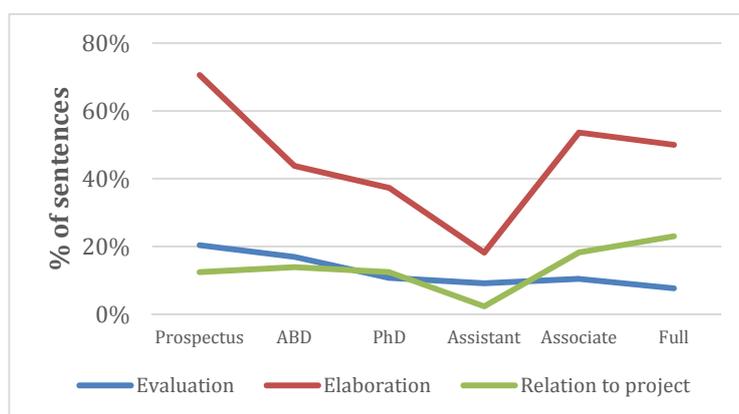
Figure 6: Comparison of participant patterns of citation use by disciplinary status.

The percentage of sentences which included some form of Evaluation shows a gradual decrease from advisees in early stages to professors with more than 10 years of experience as faculty (from a 20.3% to almost 7.7%). On the other hand, Elaboration shows the biggest change by participation level, starting at a little over 70.5%, making a substantial gradual dip until the Assistant professor level, where it drops to almost 18.2% and then increasing again for more advanced faculty, reaching around 50%. Finally, the percentage of sentences which included an explicit reference to the project was overall quite low, but showed a very substantial increase for faculty in the Associate and Full professor categories (18.2% and 23% respectively), when in the other categories it had never exceeded 13%.

In the next section of the results we present some comparisons by disciplines, collapsing the advisor/advisee pairs and range of participation distinctions.

3.2 Analysis for patterns in the disciplines

The analysis of disciplinary differences for Reference, Elaboration, Evaluation, and Relation to Project yielded some interesting, though not completely unexpected results. In terms of Reference, the ranked ANOVA test revealed a strong statistically significant relationship between disciplines and Reference with $p=.0147$. As shown in Figure 11, the four participants in Humanities and Social Sciences had on average the highest

percentage of sentences that made reference to a source (35.5%%), while the four participants in Computer Science only had an average of 13.6%. The two other disciplines, both from Engineering, fell in between and were quite close to each other, with 26.2% of the Materials Science corpus sentences making reference to a source, compared to 23.6% of the Chemical Engineering one.
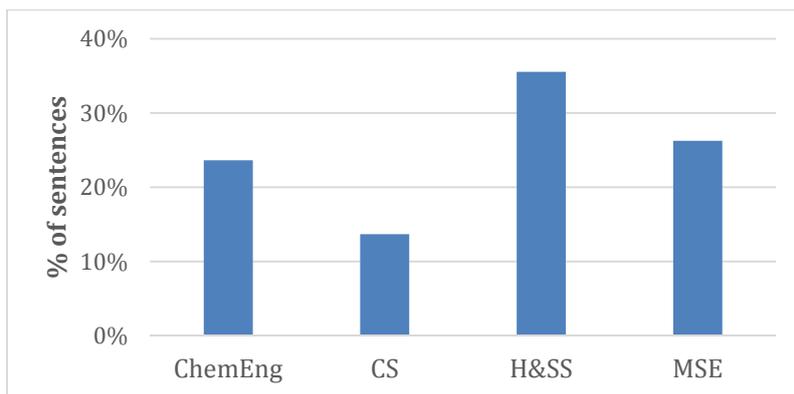
Figure 71: Overall references by discipline.

However, when examining advisors and students separately, the results are more complicated: students in H&SS and Chemical Engineering seem to be citing sources in a greater percentage of their sentences than the advisors in their discipline on average (13% more in both cases), while the opposite seems to be the case for advisees in Materials Science who reference sources by 3% less than their advisors, and Computer Science where the difference goes up to 7% less references (see Figure 12).
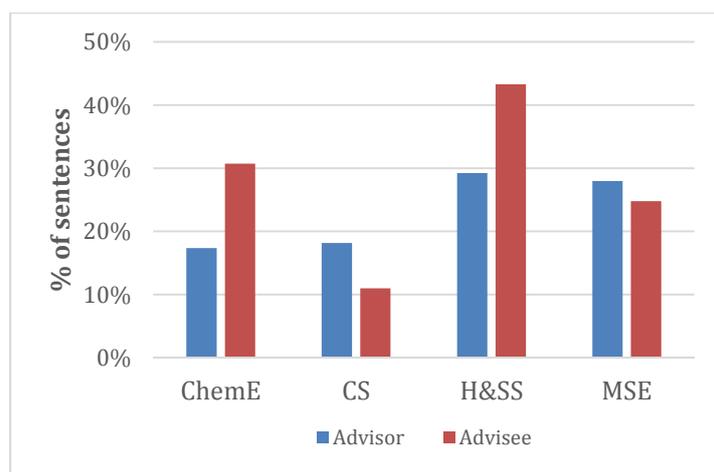
Figure 8: Comparison of references by advisor/advisees for all disciplines.

Regarding the other three categories of analysis, Figure 13 presents a cumulative picture from all four disciplines. The Ranked ANOVA tests showed significant differences for discipline for Elaboration (p=.0123), but not for Evaluation or Relation to Project.

In terms of Elaboration, Humanities and Social Sciences texts by both advisors and advisees had the highest percentage of Elaboration (almost 52%) of the sentences coded as Reference. The other three disciplines used significantly fewer elaborations, with 45% of sentences including a reference coded as Elaboration.

The dimension Relation to Project had the second highest percentage of sentences for all disciplines, except H&SS. In this case, H&SS texts had the fewest statements (9%) that explicitly showed the connection between a source and the project at hand. For all the other three disciplines, the percentage was very similar, ranging from 16.1% to 16.5%. These differences did not reach the level of significance.

Finally, the dimension of Evaluation had the overall lowest percentage of occurrences, ranging from a little over 11% (H&SS) to 14.6% (Chemical Engineering. Computer Science, in this case, was higher than Materials Science Engineering. Again, these differences did not reach the level of significance.
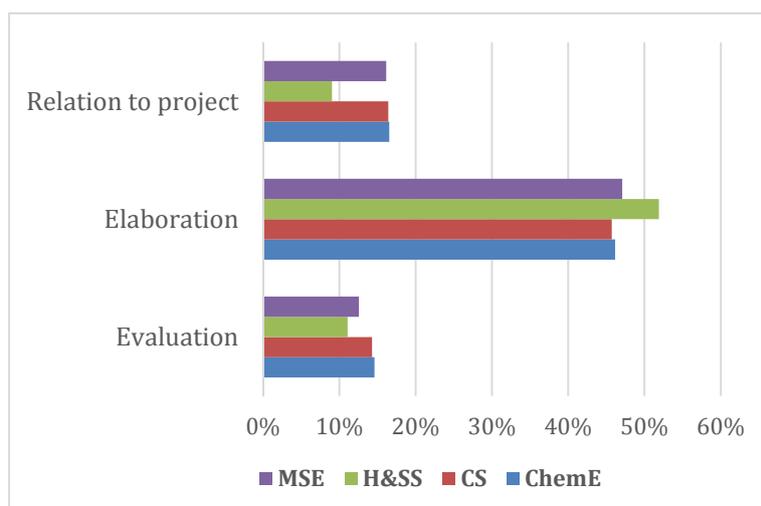
Figure 13: Comparison of reference and elaboration by discipline.

3.3 Analysis for patterns in Genres

As explained in the methodology, an effort was made to procure and analyze at least two published pieces from every participant. All advisors allowed access to their published journal articles, and one H&SS advisor even offered draft chapters of a book he was working on. However, only one instance of a book or book chapter genre would not have been sufficient for comparison purposes, so for advisors only journal articles were used for analysis (a total of 17). On the contrary, advisees were asked, if

possible, to submit a diverse range of genres, which added to 5 dissertation chapters, 2 candidacy proposals and 7 published journal articles in which they were the first authors.

Figure 14 was designed to show the combined results for all genres in the study, separating the 7 advisee published journal articles from the 17 advisor published journal articles. However, the results for journal articles were almost identical between the two groups, which explains why the two lines (green and grey) are almost identical. There was only a difference of 1% in the categories of Evaluation and Elaboration between the two categories.
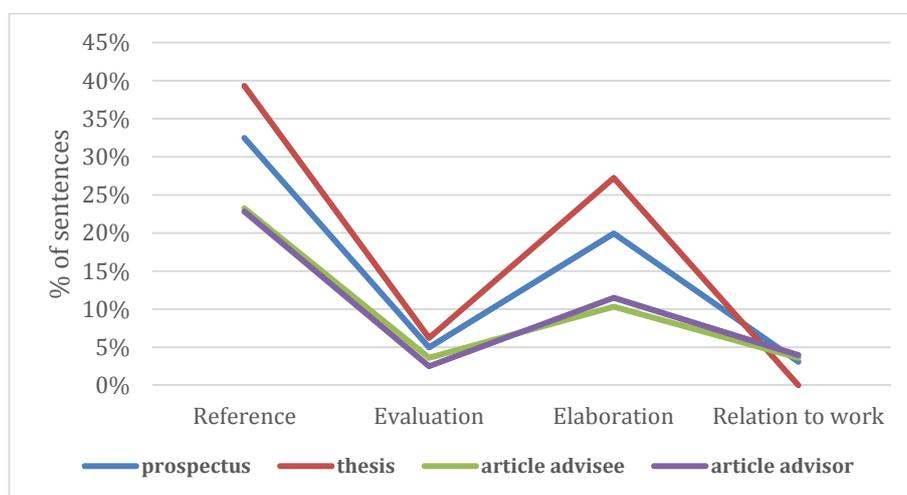
Figure 9: Comparison of genres for advisees and advisors.

Putting the advisor-authored articles aside, advisee-authored journal articles have the fewest instances of reference per overall number of sentences (23%), as well as fewest elaborations (10%) and evaluations (4%) compared to dissertations and candidacy proposals. However, the category Relation to Project was present much more in journal articles (4%) than both candidacy proposals and dissertations (3% and almost 0% respectively). On the other hand, dissertations contained over 27% elaborations from different sources, almost three times more than journal articles, and the highest percentage of references (39%), almost twice as many as journal articles.

Candidacy proposals, on the other hand, seemed to occupy a space between journal articles and dissertations, with 32% in references, 20% Elaboration and percentages closer to the journal article for Evaluation and Relation to Project.

One final point regarding the analysis of genres is that the variance between the three types of documents collected from both advisees and advisors was a result of both single and multiple source patterns. Figure 15 shows only the results from single source references and the pattern is almost identical to the one for overall references, with the

exception of slightly lower percentages in all four dimensions in journal articles authored by advisees.
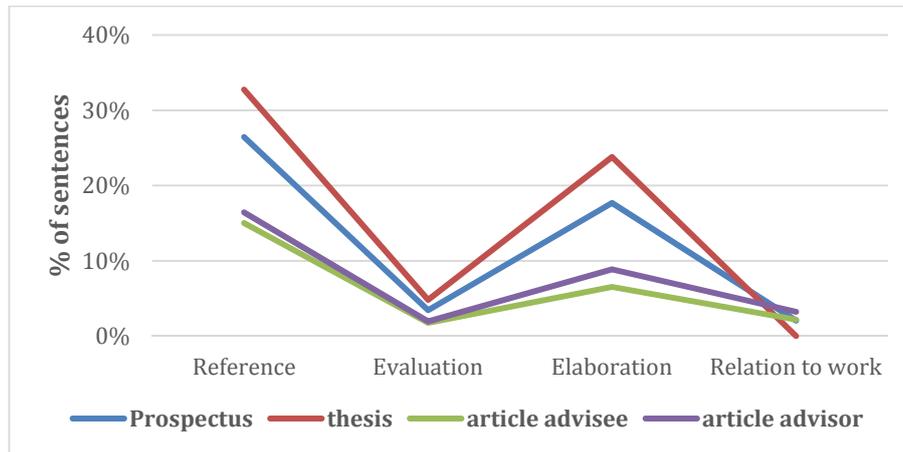
Figure 15: Comparison single source references for genres.

In conclusion, the analysis of the texts that my participants were generous enough to provide me with points to some clear patterns for disciplinarity, genre expectations, and development of levels of participation in the field. Although this is not an extended analysis of large corpora in many disciplines, identifying a number of recurring patterns may help writing researchers the processes and contexts within which authors use sources.

4. Discussion and Conclusions

Before exploring some of the possible implications of the results of this analysis, it is important to discuss the complications and effects of using a hand coding methodology for writing research. As we previously mentioned, the small size of the corpus presents a major limitation for this study, but it also made it possible to be able to prepare all the texts for coding (OCR, clean up and segmentation), and then code all the texts in the four dimensions within a reasonable amount of time. In addition, it would have been difficult to secure the help of second coders with a much larger corpus including more disciplines or genres, as the training alone would have taken much longer.

However, hand coding using the model we employed has some important benefits: first, we were able to code for multiple dimensions in several passes without a significant tax on the coders' time. Given the way the texts were organized in Excel files in sequential form, coders were able to read as if they were reading the complete article or thesis, find an instance of a reference to source, and then make all the other decisions on the other dimensions at that moment. Such a model also allowed for the

flexibility to be able to code for Elaboration in adjacent sentences to the one that actually included the reference, which is not easy to do with automated approaches.

The truth is that hand coding is a very powerful systematic approach to analyzing texts, but the cost of creating the system is quite large, between turning the data into text that can enter into spreadsheets, creating the conditions in those sheets that would minimize coding errors and misunderstandings, and finally keeping track of all the documents that are created so that they can later be analyzed. For that final stage, file naming conventions are critical, and ideally they have to be decided upon before any documents are created.

A final point about the value of hand coding comes from the ability of good human coders to provide insights about the coding categories that the coding definitions and examples had not originally accounted for. Such insights, which are often triggered by some connection or analogy to another, perhaps unrelated, system, are extremely useful not only in refining the dimensions, as was the case with this study, but also revising some of the larger research questions. In addition, human coders can not only make inferences from the context, but also from other texts they have read or coded, so there is the possibility of having better internal consistency of coding across texts – especially when the coding decisions are not based on explicit discourse. For the purposes of this project, codes based on explicit markers (such as if the author's name is present) were not favored, as we eventually wanted to be able to ask questions about the underlying motivations behind source use, not about disciplinary formalisms, so hand coding was a great option.

Keeping that larger goal of preparing for discourse-based interviews in mind, we can now turn to the discussion of some of the results of the analysis itself to explore the implications of the patterns we identified earlier.

First, in terms of the patterns by levels of participation, we found that advisees had, for the most part, more instances of Reference, Evaluation and Elaboration than their advisors. None of these differences were big enough to yield statistical significance though the patterns are suggestive. In terms of variation, we do have to take into account that advisee texts were not all published journal articles, but also candidacy proposals and dissertation chapters, so not all of the difference in advisor/advisee pairs should be attributed to the level of participation. In addition, we have to acknowledge that the difference between levels of participation was not equal across pairs or disciplines. In Chemical Engineering, both advisors were full professors, while their advisees were in early dissertation stages (one having just completed their candidacy proposal). On the other hand, in Materials Science Engineering, one pair was comprised of a new assistant professor, as the advisor, having taken over as thesis chair from a professor who had moved to a different institution, and an advisee who had significant industry experience and had completed the writing of the dissertation.

One more interesting note is related to the higher percentage of advisor sentences coded for the category of Relation to Project compared to the advisees. The genres of the dissertation and the prospectus (candidacy proposal) would probably call for such

explicit relationships between the project and previous sources to be formed. The fact that advises make this move less often might simply mean that it is more difficult for relative novices or that advisees do not always have the discourse tools to explicitly position their own project in relation to the work of others. In fact, given the differences in the sub-category of building on versus differing from a source, it seems that advisors use much more the "build from" move, perhaps being much more aware of their audience of peer-reviewers.

Another pattern in terms of level of participation can be seen when examining the range of participation, from advisees in the candidacy proposal stage to full professors (Figure 10). While it seems reasonable that advisees would be inclined to use more Elaboration and Evaluation in early stages of their career, as they often tend to summarize from their sources (as advisors often complain), it seems counter-intuitive that the trend would be reversed for advisors, with Assistant Professors being the ones who use the fewest citation patterns. However, this trend might be explained by the degree to which the more advanced advisors understand the field and are familiar with all the literature surrounding their project. It would be interesting to examine if the Assistant Professors do not have a good overview of the field or do not feel the need to include all of the sources they are familiar with. At this point we may begin to question the position of the "neophyte" in a field in terms of the need to cite more sources and use more patterns of relating them to one another and one's work. Simply examining the years of participation in a linear manner does not seem to be sufficient, at least not for the sixteen participants in the study. Thinking in terms of Zones of Proximal Development, it may be that the "official" entry into a field is also accompanied with a slight "bump" to another level of participation, where one has to establish a position in relation to the work of others almost from the beginning.

In terms of patterns in disciplines, there were some significant differences between Humanities and Social Sciences and Computer Science, with the two Engineering disciplines looking more like one another. We do know that in many Computer Science fields there is an increased emphasis in presenting the innovative solution itself (especially in published conference proceedings) instead of a more comprehensive view of the work done previously on the issue. On the other hand, in H&SS positioning a new idea in relation to previous work is extremely important, as the innovation might not be physical or technical, but conceptual. Especially when it came to Computer Science advisees, it was clear that their patterns of use were very different to H&SS advisees, who used almost four times as many references to sources than their Computer Science colleagues (Figure 12). This difference may be attributed to the reduced emphasis that computer science students seem to place on previous work, especially as the developments in the field are very fast and innovation is celebrated more than tradition.

Regarding the other three coding dimensions in the disciplines, Humanities and Social Sciences seemed to have slightly different patterns than the other three fields: in Elaboration, it was quite a bit higher, by 4-6% overall, while in Relation to Project it

was the lowest of all, with 7% average fewer instances. It seems that there might be a pattern of disciplinary understanding of what constitutes explicit relationship between a source and one's project, as the fact that one elaborates on a source might be considered a marker of increased importance for the project, at least for authors in H&SS.

Finally, the results from the analysis of genres raise the question of exigence and the demands of the audience for grounding in sources for the candidacy proposals and the dissertation. It was clear that the differences between journal article patterns overall were minimal between advisors and advisees, which could also be a result of co-authorship or vetting of the advisee published work, but the difference compared to the other genres probably cannot be accounted for by the limitation of a small corpus. In a candidacy proposal, a dissertation committee might expect an advisee to show that his/her proposal is solid by showing how it builds upon previous work — or at least how it relates to it — in order to show that it offers a unique contribution. To do this, an advisee may well incorporate more and more elaborated sources in order to establish his/her ethos as someone who is working within the research questions of the discipline. In dissertations, the percentages are significantly lower, perhaps because the research itself drives the reference to sources, so advisees are more likely to make generalized statements.

Returning to the advisee genres, we can make a couple of interesting observations especially for dissertations and articles. First, in all five dissertation chapters that we analyzed there was almost complete absence (only 0.53% for H&SS) of the category of Relation to Project. Although one would expect an increased concern by advisees in tying the citations to the argument at hand, this was obviously not the case. Another brief observation for dissertations would be that although H&SS has a much higher percentage of Reference and Elaboration, it has the lowest percentage of Evaluation, perhaps in an attempt to approximate a more "objective" model of citation.

Overall, this analysis might have not provided any concrete conclusions about the nature of citation patterns in the disciplines or across levels of participation and genres, mainly because of the small corpus size and the limited points of comparison, but it can serve well as a starting point to examine more closely some of the patterns that seemed to emerge. More importantly, however, it can serve well as a model of the ways hand coding of texts can provide a number of insights by exposing patterns and showing relationships that only human coders working within a rich context are able to uncover.

References

Ackerman, J. M. (1991). Reading, writing, and knowing: The role of disciplinary knowledge in comprehension and composing. *Research in the Teaching of English*, 133-178.

Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science* (p. 59). Madison: University of Wisconsin Press.

Belcher, D. (1994). The apprenticeship approach to advanced academic literacy: Graduate students and their mentors. *English for Specific Purposes*, *13*(1), 23-34. http://dx.doi.org/ 10.1016/0889-4906(94)90022-1

Berkenkotter, C., & Huckin, T. N. (1995). *Genre knowledge in disciplinary communication: Cognition/culture/power*. Lawrence Erlbaum Associates, Inc.

Blakeslee, A. M. (1997). Activity, Context, Interaction, and Authority Learning to Write Scientific Papers In Situ. *Journal of business and technical communication*, *11*(2), 125-169. http://dx.doi.org/10.1177/1050651997011002001

Brett, P. (1994). A genre analysis of the results section of sociology articles. *English for Specific Purposes*, *13*(1), 47-59. http://dx.doi.org/10.1016/0889-4906(94)90024-8

Brown, J. S., & Duguid, P. (1996). The Social Life of Documents; introduction by Esther Dyson. *First Monday*, *1*(1). http://dx.doi.org/10.5210/fm.v1i1.466

Bunton, D. (2005). The structure of PhD conclusion chapters. *Journal of English for Academic Purposes*, *4*(3), 207-224. http://dx.doi.org/10.1016/j.jeap.2005.03.004

Coaldrake, P., & Stedman, L. (1999) *Academic work in the twenty-first century: Changing roles and policies*. Canberra: Department of Education, Training and Youth Affairs Higher Education Division.

Connor, U. M., & Kramer, M. G. (1995). Writing from sources: Case studies of graduate students in business management. *Academic writing in a second language: Essays on research and pedagogy*, 155-182.

Cozzens, S. E. (1981). Taking the measure of science: A review of citation theories.

Cronin, B. (1984). The citation process. The role and significance of citations in scientific communication. *London: Taylor Graham, 1984*, *1*.

Culnan, M.J. (1987), ``Mapping the intellectual structure of MIS, 1980-1985: a co-citation analysis'', MIS Quarterly, Vol. 11, pp. 341-53. http://dx.doi.org/10.2307/248680

Ding, Y., Chowdhury, G., & Foo, S. (1999). Mapping the intellectual structure of information retrieval studies: an author co-citation analysis, 1987-1997. *Journal of information science*, *25*(1), 67-78. http://dx.doi.org/10.1177/016555159902500107

Dubois, B. L. (1997). *The biomedical discussion section in context* (No. 46). Greenwood Publishing Group.

Backhouse, R., Dudley-Evans, T., & Henderson, W. (1993). Exploring the language and rhetoric of economics. *Economics and language*, 1-20.

Garfield, E. (1972, November). Citation analysis as a tool in journal evaluation. *American Association for the Advancement of Science.*

Geisler, C. (1994). Literacy and expertise in the academy. *Language and Learning across the Disciplines*, *1*(1), 35-57.

Geisler, C. (2003). When management becomes personal: An activity-theoretic analysis of Palm technologies. *Writing selves/writing societies: Research from activity perspectives*. Fort Collins, CO: WAC Clearinghouse.

Geisler, C. (2011). IText revisited: The continuing interaction of information technology and text. *Journal of Business and Technical Communication*, *25*(3), 251-255. http://dx.doi.org/ 10.1177/1050651911400701

Gibbons, R. (1998). *Incentives in organizations* (No. w6695). National Bureau of Economic Research

Gmür, M. (2003). Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, *57*(1), 27-57. http://dx.doi.org/10.1023/A:1023619503005

Harwood, N. (2005). 'Nowhere has anyone attempted… In this article I aim to do just that': A corpus-based study of self-promotional I and we in academic writing across four disciplines. *Journal of Pragmatics*, *37*(8), 1207-1231. http://dx.doi.org/10.1016/j.pragma. 2005.01.012

Harwood, N. (2009). An interview-based study of the functions of citations in academic writing across two disciplines. *Journal of Pragmatics*, *41*(3), 497-518. http://dx.doi.org/10.1016/ j.pragma.2008.06.001

Hunston, S. (1993). Professional conflict: Disagreement in academic discourse. In M. Baker, G. Francis and E. Tognini-Bognelli (Eds.), Text and Technology: In Hotiotir of, John Siticlair (pp. 115-134). Amsterdam: Benijamins.

Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied linguistics*, *20*(3), 341-367. http://dx.doi.org/10.1093/applin/20.3.341

Hyland, K. (2001). Humble servants of the discipline? Self-mention in research articles. *English for specific purposes*, *20*(3), 207-226. http://dx.doi.org/10.1016/S0889-4906(00)00012-0

Hyland, K. (2002). Authority and invisibility: Authorial identity in academic writing. *Journal of pragmatics*, *34*(8), 1091-1112. http://dx.doi.org/10.1016/S0378-2166(02)00035-8

Kärki, R. (1996). Searching for bridges between disciplines: an author co-citation analysis on the research into scholarly communication. *Journal of information science*, *22*(5), 323-334.

King, D. W., McDonald, D. D., & Roderer, N. K. (1981). *Scientific journals in the United States*. Dowden, Hutchinson & Ross; distributed by Academic Press.

Leydesdorff, L. (1998). Theories of citation?. *Scientometrics*, *43*(1), 5-25. http://dx.doi.org/10.1007/BF02458391

Leydesdorff, L., & Amsterdamska, O. (1990). Dimensions of citation analysis. *Science, Technology & Human Values*, *15*(3), 305-335. http://dx.doi.org/10.1177/016224399001500303

Martens, B. V. D. V. (2001). Do citation systems represent theories of truth. *Information Research*, *6*(2), 6-2.

Martyn, J. (1965, June). An examination of citation indexes. In *Aslib proceedings* (Vol. 17, No. 6, pp. 184-196). MCB UP Ltd. http://dx.doi.org/10.1108/eb050021

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science (1986-1998)*, *41*(6), 433. http://dx.doi.org/10.1002/(SICI)1097-4571(199009)41:6<433::AID-ASI11>3.0.CO;2-Q

McGinley, W. (1992). The role of reading and writing while composing from sources. *Reading Research Quarterly*, 227-248. http://dx.doi.org/10.2307/747793

Myers, G. A. (1992). Textbooks and the sociology of scientific knowledge. *English for Specific purposes*, *11*(1), 3-17. http://dx.doi.org/10.1016/0889-4906(92)90003-S

Nielsen, S. M., & Rocco, T. S. (2002). Joining the Conversation: Graduate Students' Perceptions of Writing for Publication.

Odell, L., & Goswami, D. (Eds.). (1985). *Writing in nonacademic settings*. Philadelphia PA: Guilford Press.

Paul, D. (2000). In Citing Chaos A Study of the Rhetorical Use of Citations. *Journal of Business and Technical Communication*, *14*(2), 185-222. http://dx.doi.org/10.1177/105065190001400202

Prior, P. (1998). Writing/disciplinarity. *Mahweh, NY Lawrence Erlbaum*.

Spivey, N. N. (1984). *Discourse synthesis: Constructing texts in reading and writing*. S. J. Samuels (Ed.). International Reading Association.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press. http://dx.doi.org/10.1177/0741088387004002004

Swales, J., & Najjar, H. (1987). The writing of research article introductions. *Written communication*, *4*(2), 175-191.

Thompson, P., & Tribble, C. (2001). Looking at citations: Using corpora in English for academic purposes. *Language Learning and Technology*, *5*(3), 91-105.

Wang, P., and Soergel, D. (1996). Beyond topical relevance: Document selection behavior of real users of IR systems. In Proceedings of the 59 th Annual Meeting of the American Society of Information Science, 87-92.

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for information Science*, *32*(3), 163-171. http://dx.doi.org/10.1002/asi.4630320302

Appendix: Description of coding categories

Reference

- *Single source*: any sentence where the author references one source directly or indirectly by
  - *Name/title/category only (no citation)*: any reference to a single source where the author only mentions a name (e.g. "Freud said") or a title (e.g. "the director of Bell Labs believes") or a general category (e.g. "a philosopher once said") and there is NO citation pointing to a specific work where the information or argument can be found.
  - *Parenthetical citation or brackets or footnote number only*: any reference to a single source where the author provides a parenthetical citation (e.g. (Vygotsky 1967)) or a bracket (e.g. [4]) or a footnote (e.g. "writing[7]") pointing to a source.
  - *Name and parenthetical citation/brackets/footnote*: any reference to a single source where the author's name is part of the sentence and is followed by a parenthetical citation/brackets/footnote (e.g. "Vygotsky (1967) argued" or "Shaw has shown that … [5]).
  - *Name, accompanied with details (and citation):* any reference to a single source where the author's name is part of the sentence and is followed by some details about the work or the person being cited as well as a parenthetical citation, brackets or footnote (e.g. "Vygotsky, in his seminal work *Mind in Action*, (1967) argued" or "Shaw, one of the most significant researchers in ABC, has shown that… [5]).

- *Multiple sources:* any sentence where the author references many sources directly or indirectly by
  - Names/titles/category only (no citation): any reference to many sources in a single sentence where the author only mentions names of authors (e.g. "Freud and Lacan said") or titles (e.g. "the board of directors of JAMA believes") or a general category (e.g. "experimentalists have argued") and there is NO citation pointing to a specific work where the information/argument can be found.
  - Parenthetical citation or brackets or footnote number only: any reference to many sources in a single sentence where the author provides a parenthetical citation (e.g. (Vygotsky 1967, Engeström 1993)) or a bracket (e.g. [4, 5]) or a footnote (e.g. "writing7,8") pointing to the sources.
  - Names and parenthetical citation/brackets/footnote: any reference to many sources in a single sentence where the authors' names are part of the sentence and are followed by a parenthetical citation/brackets/footnote (e.g. "Vygotsky (1967) and Engeström (1993) have argued" or "Shaw [5] and Smith [6] have

shown XYZ). Do NOT code for 3 if the reference is for a single source with multiple authors (e.g. "Shaw and Smith [5] have shown.")

- *Names, accompanied with details (and citation)*: any reference to multiple sources where the authors' names are part of the sentence and are followed by some details about the work or the person being cited as well as a parenthetical citation, brackets or footnote. Also any reference to many sources in a single sentence where one source is presented separately from the others (e.g. "Vygotsky (1967) argued XYZ (see also Bakhtin 1988 and Emerson 2001)."

Evaluation

- *Single source evaluation*: any sentence where the author makes an evaluative statement or comment about a source EITHER by
  - noting the (established) value of the source/idea, where the author points to the usefulness or positive impact of the source for the field, the specific project or the understanding of a new concept through phrases such as "another significant research was conducted by…" or "this methodology [4] provides the foundation/standard for…" or "this has been used extensively [5]" or "widely used" or verbs such as "the authors point out that"

  OR by
  - noting the limitations of the source/idea, where the author points to the problematic nature or limitations of the source for the field, the specific project or the understanding of a new concept through phrases such as "this position [5] fails to take into account…" or "Smith's argument [6] is based on the assumption that…, which has proven to be wrong" or "this method [11] presents difficulties in determining…"

- *Multiple source evaluation*: any sentence where the author makes an evaluative statement or comment about many sources EITHER by
  - noting the (established) value of the sources/ideas, where the author points to the usefulness or positive value of these sources for the field, the specific project or the understanding of a new concept through phrases such as "significant research was conducted by [3], [4], [5]" or "the work of Freud (1922), Lacan (1982) and Derrida (1988) has provided the foundation for…"

  OR by
  - noting the limitations of sources/ideas, where the author points to the problematic nature or limitations of the sources for the field, the specific project or the understanding of a new concept through phrases such as "experimentalists [4], [5] have failed to take into account…" or "this common argument [6, 7, 11] is based on the assumption that, which has proven to be wrong" or "this method [11,12] shows difficulty in determining…"

  OR by

- showing how one source is better/improves upon or is inferior to others, where the author describes the relation between the sources in a positive or negative way through phrases such as "extending the work of [12, 18 and 22], Newman et al. provided an innovative solution to the problem of…" or "this is a common argument among theorists [6, 7, 11], although [11] does not apply to…"

Elaboration (one or subsequent sentences)

- *Single source elaboration*: any sentence where the author elaborates on the information or ideas being presented from a cited source, even if the reference to this source appears in a previous or later sentence EITHER by
  - providing details on specifics from the source, where the author presents specific details on the information or concepts presented in the cited source by explaining a series of results, or describing the source's whole line of argument through phrases such as "In a previous paper [7] we discussed XYZ, and we developed a framework where…" or "changes due to the nature of the materials have been reported [7]"

  OR by
  - relating the source to a general category/school of thought, where the author relates the information or concepts presented in the cited source to a general category or field of inquiry or theoretical approach through phrases such as "Positivists such as Ayer (1987) have argued that…" or "Similar progress in aligned carbon nanotubes has been made [2]" or "Scholars have long criticized the idea that…"

- *multiple source elaboration*: any sentence where there is elaboration on the information or ideas being presented from more than one cited sources, even if the reference to this source appears in a previous or later sentence EITHER by
  - noting areas of agreement between sources, where the author establishes a connection between different sources to show how they present similar information or they agree in their approach, methods or results through phrases such as "Lave (1993) and Wenger (1997) take a similar approach" or "From Lev Vygotsky through Yrjö Engeström to PD practitioners such as Suzanne Bødker, activity theory has emphasized…" or "nanoscale fillers have attracted interest because… [8-11]

  OR by
  - noting conflict, disagreement between sources, where the author establishes a connection between the sources to show how they disagree or one presents a different perspective than the others through phrases such as "Results have been consistent in the literature [1-5] with the exception of [6] where…" or "In contrast to our approach introduced first in [13], this model…"

Relation to current project (one or subsequent sentences)

▪ *Note that own project builds on existing work(s):* any sentence where the author makes an explicit statement to show how the current project builds upon or is an extension of a previous project from a cited source. Examples: "the methodology that we used has been well-established in [4,5] or "following the work of Freud, I want to argue that…" or "LDPE [12] was chosen as a representative polymer matrix".

Note that own project differs from existing work: any sentence where the author makes an explicit statement to show how the current project is different from a previous project which is referenced. Examples: "this paper tries to overcome some of the limitations of Smith (2001)" or "In contrast to our approach, introduced first in [13], this model…."