Analyzing the Language of Citation across Discipline and Experience Levels: An Automated Dictionary Approach

David Kaufer, Suguru Ishizaki & Xizhen Cai

Carnegy Mellon University, Pittsburgh | USA

Abstract: Citation practices have been and continue to be a concentrated area of research activity among writing researchers, spanning many disciplines. This research presents a re-analysis of a common data set contributed by Karatsolis (this issue), which focused on the citation practices of 8 PhD advisors and 8 PhD advisees across four disciplines. Our purpose in this paper is to show what automated dictionary methods can uncover on the same data based on a text analysis and visualization environment we have been developing over many years. The results of our analysis suggest that, although automatic dictionary methods cannot reproduce the fine granularity of interpretative coding schemes designed for human coders, it can find significant non-adjacent patterns distributed across a text or corpus that will likely elude the analyst relying solely on serial reading. We report on the discovery of several of these patterns that we believe complement Karatsolis' original analysis and extend the citation literature at large. We conclude the paper by reviewing some of the advantages and limits of dictionary approaches to textual analysis, as well as debunking some common misconceptions against them.

Keywords: citation research, common archives, corpus analysis, dictionary methods, text analysis

journal of WRITING RESEARCH

Kaufer D., Ishizaki S., & Cai X. (2016). Analyzing the language of citation across discipline and experience levels: An automated dictionary approach. *Journal of Writing Research*, *7(3)*, 453-483. doi: 10.17239/jowr-2016.07.03.07

Contact: David Kaufer, Carnegie Mellon University, Department of English, Baker Hall 145 F, 5000 Forbes Avenue Pittsburgh, PA 15213 | USA - kaufer@andrew.cmu.edu

1. Introduction

The present study reanalyzes the data from Karatsolis' research on citation (2005), which was focused on comparing the citation behaviors of 8 PhD advisees and their 8 advisors from four disciplines (Chemical Engineering [CHEME], Humanities and Social Sciences [HSS], Materials Science Engineering [MSE], Computer Science [CS]). Our interest was not to replicate or call into question Karatsolis' coding results or findings. Instead, in the spirit of this special section on the promise of reanalyzing a common dataset in writing research using complementary methods (see Cheryl Geisler's introduction), our contribution focuses on showing what so-called "dictionary" approaches can bring to the study of citation.

Our general theoretical framework, operationalized in the automatic coding environment known as the DocuScope Text Analysis Environment (Ishizaki & Kaufer, 2011), was designed to help text analysts investigate rhetorical variation in a single text or textual archive across 31 "discourse-wide" clusters of rhetorical variables (see appendix 1). Studying citation practices is a focus of one of our pre-existing clusters, which we adapted from the CARS (Create a Research Space) model of Swales (1990).

2. Our Theoretical Approach to Citation

Our approach breaks down citation into three components: (1) Selecting the language of secondhand authority, which opens the option of citation; (2) The decision to make a citation once committed to that selection; (3) The visible format of a citation once the decision to cite has been made. Of the three components— the selection of citable language, the decision-to-cite, the formatted citation—our theoretical framework focuses principally on the first phase, though it can be used to investigate the other phases as well. It should be noted that the decision-to-cite, and more specifically the need to supplement one's own knowledge with secondhand sources, has recently been empirically shown fundamental to the writing processes underlying professional communications in the workplace. Leijten, Van Waes, Shriver and Hayes (2014) recently updated Hayes's 2012 process model of writing to accommodate how writers search and select task-relevant external digital sources to perform workplace writing tasks. Looking outward from the searching and selection of sources, older citation research has focused on the cognitive (Price, 1963; Weinstock 1971; Cronin 1984; McInnis & Symes 1989), social (Moravsik & Murugesan 1975) persuasive (Gilbert 1976; 1977), tactical (MacRoberts & MacRoberts, 1986) and rhetorical (Small, 1978; Kaufer & Carley 1993: chapter 8; Paul, 2000) underpinnings of citation. Beyond these studies, there has been a recent wave of scholarship exploring the "cultures" of intellectual property and in particular how postmodern theories of authorship, the digitization of writing, and the corporate strengthening of copyright have jointly contributed to the complication of these cultures both for cultural producers and those seeking to make

fair use of their work (see the special issue on writing and western conceptions of intellectual property edited by Kennedy and Howard, 2013).

Within our theoretical perspective, and the perspective of many researchers (Ritter, 2005; Lunsford, Fishman, & Liew, 2013; Jamieson & Howard, 2013), textual citation does not stop with linking an idea to a particular bibliographic entry to assert intellectual influence and to acknowledge another's intellectual property. Citation performs a range of additional cognitive and social functions discussed in the literature: from supporting a specific argument, conveying a map of the field, projecting research identities (Hyland, 2012), aligning with and debunking camps, and many more (MacRoberts and MacRoberts, 1986; Harwood, 2009). Before student writers can be taught to cite, they must first learn how to write in an academic register featuring reference to secondhand authority. Students must learn, mostly implicitly, that academic writing proceeds on the strength of authority embodied in Newton's aphorism that "If I have seen farther, it is by standing on the shoulders of giants." Robert Merton, one of the founders of the sociology of science argued that establishing priority (viz., symbolic intellectual property) through secondhand authority is a central motivator of scientific discovery and that citation provides cultural validation that priority has been successfully established (Merton, 1957; Stephan, 2004).

Still, the linguistic and discourse characteristics of citable language remain an important and understudied challenge in citation studies. By "citable language," we mean the lexicogrammatical structures and functions that readers in a discipline must learn in order to discern an author's use of secondhand authority; and the lexicogrammatical structures and functions that writers in a discipline must internalize in their own practice if they are to convey to readers their reliance on secondhand authority. Plagiarism resources like Purdue's OWL[1] and the Council of Writing Program Administrators statement on plagiarism[2] provide excellent conceptual overviews of plagiarism and plagiarism's encroachment on the intellectual property of secondhand authorities. But even these vaunted sources are silent about the manifold structures and functions of English that can convey secondhand authority in English. Writing Centers that seek to be highly "hands-on" for students about avoiding plagiarism, like Southwestern University's Debby Ellis Writing Center, advise students as follows: "Any time you use *words, ideas, data, images, or theories that are not your own* (author's emphasis), you need to let your reader know who did the work of figuring (or spelling) these out, and where we can find these phrases or images, this data, or these ideas." This language commendably overviews the range of objects of secondhand authority needed to trigger citation. But it leaves undeveloped an operational understanding of the manifold linguistic embodiments that secondhand authority can take on the page. And such operationalization is essential for advancing the subfields of writing research that deal with citation practices.[3]

What makes the study of citation and these linguistic embodiments well-suited for dictionary approaches to text analysis is the fact that expressions conveying secondhand authority in English hew to well-recognized patterns that can be

enumerated in a rule-governed way. Swales (1990) and his collaborators have studied how linguistic structures associated with citation fit within discourse functions that Swales calls "rhetorical moves." In fact, through his pedagogical and corpus work, Swales' findings can be viewed as providing small training dictionaries linking sentences of English with rhetorical moves for pedagogical purposes. So, for example, Swales classifies sentences of the form "There has long been interest in X" as a writer's effort to establish an ongoing interest in a niche research area. And he classifies sentences like, "Since the 1990s, there have been a spate of studies on X" and "Jones found that X" as a writer's effort to review previous research in that particular area. Swales' work was instrumental in seeding the DocuScope dictionaries to cover academic citation. However, the DocuScope environment made it possible to scale Swales' patterns to tens of thousands of full or partial sentence patterns associated with citation. We call these patterns our "citation-specific" patterns (see entry 2 in appendix 1 and appendix 2). But machine learning approaches, like those utilized by Bill Hart-Davidson and Ryan Omizo in this volume, can also be used to capture the lexicogrammatical properties of academic citation and can scale to a large number of patterns with ostensibly much less human labor costs than dictionary methods. We say "ostensibly" because the jury is still out, we believe, on the investment in human annotation required to "train" machine learning programs analyzing writing to a desired standard of accuracy. Cotos (2014) recently published a monograph-length study on a machine-learning program, the Research Writing Tutor, she and her colleagues at Iowa State developed to classify Swales-style moves across research papers (Cotos, Huffman, Link, 2015). To try to ensure the precision of the Tutor, Cotos recruited a multi-year, multi-disciplinary team that painstakingly annotated 900 research articles over 30 disciplines. Machine-learning approaches for writing scale better than dictionary approaches, but they can do so at the cost of precision. Dictionary approaches, for their part, are labor intensive and their ultimate sustainability, we suggest, in all likelihood depends on aligning them with machine-learning methods. We seek to reinforce this argument further at the end of this paper.

In order to cite, students must learn that reference to secondhand authority requires learnable linguistic constructions (Goldberg, 1995) characterizing externalized points of view (e.g. "Jones argues for;" "it is widely thought that"). These constructions fall into what Geisler has termed a "main/faulty path structure" (Geisler 1994, pp. 143-148), where an author characterizing a source in the third person cues it as a constructive point of departure for inquiry (main path) or a rejected detour (faulty path) from acceptable paths forward. The metaphor of "path" is carefully chosen because the academic register constitutes a language of directionality with pathways coursing through it. These pathways project a sophisticated social model that posits an expert class creating knowledge through reasoned and (in the sciences) replicable method and judicious interpretation and argument. The knowledge-creation process depends on following recorded paths judged fruitful and resisting paths judged barren. Students who acquire the academic register with an insider's understanding learn that to make

their own position heard, they must give an expert class "presence" (Perelman & Olbrects-Tyteca, 1969) in their lexicogrammatical choices.

Understanding academic language as pathways from secondhand authority to take or resist presents a sophisticated understanding. Learning the language of secondhand authority often precedes formal citation practice by many years and starts to insinuate itself when the textual focus of the American language arts curriculum migrates from narrative to information. For example, should American eighth-graders read a grade-appropriate information article on albino tigers and be asked to summarize it, they are expected to progress from explicit firsthand representations (e.g., "I read about albino tigers") or implicit ones (e.g., "the cause of albinoism in tigers is...") to secondhand representations that delineate an expert class as an explicit subject (e.g., "scientists have long studied albinoism in tigers) or direct object (e.g. Alibinoism has long fascinated scientists). They learn they can foreground an institutionalized cognition with an unfilled agent slot for the expert class (e.g., "Albinoism in tigers has been studied for decades"). They learn they can cast "research" as a nominalization (e.g., "Albinoism is a growing area of research") and then leave it for readers to infer the existence of a class of researchers who are agents of the research.

Sentences that represent secondhand authority require citation in principle. But in practice, learning the lexicogrammatical structures of secondhand authority and acquiring even a rudimentary understanding of citation can occur at displaced phases of instruction. In the 2015 Pennsylvania assessment for eighth grade language arts (PSSA, 2015), the rubric for "information" writing requires students to identify and formulate secondhand representations such as "the author wishes to show," "the author believes," and "the authors wants to describe" (50). The guide repeatedly reinforces the use of these references to secondhand authority as a means of extracting evidence from the text, containing 63 references to the word "evidence" alone. But tellingly, the same guide contains not a single mention of "citation," or the need to credit the author or the author's work in a "work cited" section.

Yet by the freshman year of college in America if not well before, explicit citation education rises to the top of ways students are expected to document their interaction with a published author's ideas and expression. Competing perspectives on citation, what Lunsford, Fishman and Liew (2013, p. 476) label "insider" and "outsider" perspectives, converge to explain why citation ascends to such a pivotal place in the U.S. postsecondary curriculum. From the outsider perspective, citation is framed as a corrective against plagiarism and intellectual theft. The enforcement of citation as an anti-plagiarism deterrent has a long history in western classrooms (Howard & Robillard, 2008), and the apparatus of enforcement has grown significantly in recent years with the infusion of international students who come from countries where the research paper is not part of the middle or high school curriculum, where a limited English vocabulary can make paraphrase ("put in your own words") challenging, and where close imitation without acknowledgement may not be viewed as a serious offense (Mertha, 2007).

The insider perspective brings a shift of emphasis and, more importantly, a shift of tone to the understanding of citation. It elevates citation from the deterrent of "theft prevention" to the nurture of "cherished community value to uphold." Students are encouraged to think of citing others as essential apprenticeship for eventually being cited (Kaufer and Geisler, 1989; Swales, 1990; Geisler, 1994; Lunsford, Fishman & Liew, 2013, document some students making the transition from outsider to insider over four years at Stanford). They are challenged to think of themselves as innovators-in-the-making. And in this role, they are given to understand that invention in a social context builds on mastering the social networks (Kaufer & Carley, 1993; Lunsford, Fishman & Liew, 2013), communities of practice (Wenger, 1998), task-relevant sources (Leijten, Van Waes, Schriver & Hayes, 2014) and community understandings (Dong, 1996) that can buttress and extend their own ideas. From the inside, students learn that knowing whom to cite requires a special savvy of belonging to a field and sharing its mission. They learn that publishing and earning citation for one's own work rests on respecting fragile assumptions of collegiality and shared purpose that allows "new" ideas to be assimilated without encroaching too far on existing relationships of power, authority, and status (Kaufer and Geisler, 1989; Kaufer and Carley, 1993, Hyland, 2004; Lunsford, Fishman & Liew, 2013).

Nonetheless, even as they learn to simulate the external citation behavior of insiders, students can retain a decidedly outsider orientation to citation as a general practice. As Karatsolis found (above), the largest number of citations produced by established and emerging scholars alike amounted to referential "knowledge-telling" (Scardemalia & Bereiter, 1987) strategies that align citation with showing one has done the necessary "homework" (Karatsolis, 2005, 84). In his discourse-based interviews, Karatsolis learned that advisors with years in the field had a greater grasp of sophisticated citation strategies than their advisees, findings further reinforced by Thompson and Tribble (2001), Petrié (2007), Harwood (2009) and Mansourizadeh and Ahmad (2011).

With this general background about citation, our aim in this paper is to analyze Karatsolis' corpus within our dictionary-based automated text analysis environment.

3. The DocuScope Text Analysis/Visualization Environment

The DocuScope text analysis/visualization environment has been described in previous publications (Kaufer, Ishizaki, Butler, Collins, 2004; Ishizaki & Kaufer, 2011) and applied in others (Kaufer, Ishizaki, Collins, Vlachos, 2004; Collins, Kaufer, Vlachos, Butler, Ishizaki, 2004; Kaufer, 2006; Kaufer & Hariman, 2007; Al-Malki, Kaufer, Ishizaki, Dreher, 2012). Hence, in this section, we provide a brief overview of some of the highlights of its history and features. The project began life in the late 1990s when the present authors (David Kaufer and Suguru Ishizaki) embarked to investigate ways of doing rhetorical analysis of texts through text visualization. At the time, we believed that visualization of text could augment serial reading processes to advance techniques

of critical assessment. To undertake experiments in visualization, we investigated text-processing environments already widely used at the time for automatic analysis (Hart, 2013; Pennebaker, 2012; Biber, 1989). Biber provided a useful and a still very influential functional breakdown of spoken and written English (most famously distinguishing "informative" from "interactive" language). However, his tagging categories focused mostly on modality (speech vs. writing) and high-level genre functions (information vs. story) and were not originally designed to assist in the semantic or rhetorical interpretation of particular texts.[4] The dictionaries of Hart and Pennebaker were based on more explicitly psychological and rhetorical interpretive interests respectively. Both dictionaries cover single words only and only on the order of 2,000 to 10,000 words. Such size is perfectly fine for the tasks they were designed to support—statistically clustering texts of interest thematically against a large background of reference texts. When the reference set is large enough, even small dictionaries can perform with great accuracy classifying a text of interest against the reference set. For example, Rod Hart's pioneering DICTION program can accurately gauge that a political speech of interest scores high, say, on themes specific to his system, themes like "certainty," "optimism," and "centrality" and score low, say, on the theme of "realism." DICTION can make these assessments very accurately by comparing the speech of interest against 30,000 reference texts that have been scored on the same themes. When you actually read the speech analyzed, however, you may scratch your head wondering where and how these themes unfold in the text before you. The profile of the text is an accurate score, but it is an *aggregate* score and can leave a light footprint on the text as a *particular* specimen. In a post-Bakhtinian world, we have no illusions that there exists anything like a *unique* text. Language recycles too massively across texts to support claims of uniqueness. But textual particularity still captures the important fact that individual texts have their own signatures of the language they recycle and their own distinctive frequencies of the patterns in which they recycle it. When you are doing thematic analysis, as Hart and Pennebacker tend to do, your aim is to see the texts of interest through the lens of big data and you are accomplishing your ends with a smaller investment in textual particularity. But if you are trying to give the serial reader more ammunition for interpreting a particular text, the light footprint of large aggregate samples on particular texts can leave the would-be interpreter of those texts under-supported.

These considerations led us to conceive of a text analysis tool that could "cover" textual particularity by recognizing not only single words but word sequences (strings) of any length. But where would the strings and their categorizations come from? There were no existing English-language references that archived the millions of "runs" of reusable English that help define the everyday language we draw upon effortlessly and unconsciously to make meaning. We recognized we would need to harvest these strings from the wild and create our own archive. We further recognized that the same visualization techniques we had been exploring to visualize existing dictionaries could help us build this archive, could help us develop novel functional-semantic dictionaries

for automated rhetorical analysis on a scale many times larger than previous dictionaries.

The system's ability to capture phrase and even clause-level word sequences enabled us to account for a rich repository of serendipitous semantic variation as one word transitions into another. It allowed us, for example, to classify "swear at" a negative relationship and "swear by" a positive one. It allowed us to record that if circumstances "left one high," the expression signaled a private mental state, but with the transition into "and dry" (viz., "left one high and dry"), a new semantic space opens of negative desperation. It allowed us to record that there is positive value in "holding one's own" but "holding one's own counsel" transitions into private experience. We learned that in the post-verb slot, an "oversight" (e.g., "is an oversight," "due to an oversight," "was guilty of oversight," "committed an oversight") signals insufficient attention, but in a subject NP position (e.g., "committee oversight belonged"), a direct object with certain verbs (e.g. "took oversight for") or an object of certain prepositions (e.g., "under the watchful oversight of"), the attention signaled is supervisory and authoritative.

DocuScope's suite of interactive visualizations provide dictionary-building teams a "jeweler's loupe" into troves of these hard-to-detect and hard-to-systematize semantic-transitions and made it possible for teams to notice, extract, classify, and systematically archive them on a massive scale. The first author and his colleagues and students have given these dynamic visualization interfaces a daily work-out over many years to build the 31 discourse-wide dictionary dimensions (appendix 1) that exists today. This dictionary, sometimes referred to as the "default" dictionary, now contains more than 50 million[5] uniquely classified patterns, ranging from 1 to 13 words in length, systemically organized by dimensions and subdimensions. The entire environment, along with the default dictionary, is freely available for download from Carnegie Mellon.[6] Researchers can use the default dictionary or build their own customized dictionaries using the same visualization support used to build the default dictionaries.

4. Operationalizing Citation Measures for Automated Coding

To measure citation practices for automatic coding within DocuScope, we relied both on our discourse-wide 31 dimensions (appendix 1) and the 13 citation-specific subdimensions (appendix 2). These subdimensions differ from Karatsolis' categories of citation reference, citation evaluation, and citation elaboration. For Karatsolis, citation reference is defined as "any instance where there is an explicit or implicit reference to a source, regardless of the presence or type of citation" (59). Our category of reference is similar to Karatsolis' but because we used automatic and not interpretive coding, we depended on explicit signals in the input stream for clues of author/date/page or numerical in-text citation.

For Karatsolis, citation evaluation is defined as "any instance where there is an explicit or implicit evaluation of the cited source." (59) Within our framework, the most salient evaluative binary is whether the citation has a proven authority or whether its

authority remains a "claim" lacking general acceptance (Thompson and Ye 1991). For Karatsolis, phrases like "significant research," (66), "has been used extensively," and "widely used" (67) attest to a positive evaluation of the citation. In our framework, it mainly attests to the authoritativeness of the citation. To earn a "positive" evaluation in our system, there needs to be an active and independent signal from the positive value dimension (appendix 1, dimension 32). Further, when coming across a positive evaluation in the input stream (e.g., "in his widely-cited and brilliant study"), we record the positive evaluation ("brilliant") and authoritativeness ("widely-cited") of the citation as independent judgments.

The difference between Karatsolis' approach and ours speaks to the differences in our coding environments. Karatsolis had to write directions for human coders who are robust interpreters and who can integrate judgments of valence (positive vs. negative evaluation) and judgments of authority (claimed vs. established knowledge) without even realizing they have made the integration or crossed boundaries doing so. You would need to give human coders exotic examples like "I benefited (high positive evaluation) reading his interesting but misguided (failed authority) position" to persuade them that positive evaluation of a citation does not commit one to its authority. Similar conceptual consolidations obtain in Karatsolis' instructions to coders about negative evaluation. For Karatsolis, negative evaluations of a citation appear in statements like "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" (67). In our framework, phrases like "fails to take into account" and "proven to be wrong" are coded as negative valuations (appendix 1, dimension 31), not unauthorized citations. Our dictionaries count the phrases "this position" and "Smith's argument" as independent signals of contestation, and so signal these as citations resting on claimed rather than established authority.

An interpretative coding scheme can cross boundaries and omit much with impunity because in the normal course of reading human coders are constantly supplying gap-filling inferences to make meaning. Human coders can recognize when conventional meanings are being enforced and when they are being suspended. Instruct a human coder that words like "failed" and "flawed" around a citation will signal an unauthorized citation and it seems entirely credible. The coder will find these instructions reinforced a preponderance of the time when landing on sentences like: "Position [5] is unfortunately failed and flawed." But as long as they are allowed to read while coding, human coders won't be fooled if the input text contains craftily worded passages that can undermine these instructions: "Position [5] failed to reveal any flawed thinking. The word "failed" in this sentence suspends the conventionally negative force of "flawed" to produce an implied positive evaluation. Attentive readers are not fooled by this serendipity. Unless they are built with this serendipity built in, automated coding systems are fooled badly. Our aim with coding was to combine the speed and consistency of automation with as much serendipity as we could anticipate -- fully aware that all the serendipity we could anticipate would be at most a drop in the ocean of all the serendipity a human reader learns over a lifetime of reading how to handle.

But it is essential to bear in mind that our dictionary-building efforts were never designed to replace human reading. We rather sought to give the reading brain a "third eye" against which to triangulate on the serial reading process (Hope and Witmore, 2007). And if the corpus is bounded and fixed, researchers relying on close reading and utility software can fill in a large portion of the serendipity housed in a corpus in a relatively short time.

Our framework for citation (appendix 2) covers 5 subcategories we respectively label cited authority, cited claims, cited references, cited gaps, and cited quotations. Cited authority for us assumes the cited reference is an established knowledge source in the field. As Karatsolis points out as part of his definition of positive citation, cited authority is associated with a raft of locutions such as: "As Jones demonstrated (1972)," "has long been accepted," and more. Cited claims in our framework are more agnostic and sometimes negative. They refer to proposed knowledge whose acceptance as knowledge still awaits a verdict. In our independent investigation of the language of citation, we have found that cited claims cover a very wide swath. They are presented as knowledge still contestable (e.g., "Jones has argued") or contingent (e.g., "If Jones' theory proves right"; "Jones may have discovered..."). Cited claims also include claims by the author to counter previous claims whose acceptance has been established or remains pending (e.g., "these results contradict the widely-held view"). Further, they include claims of self-citation in a current work where the author makes an implicit pitch for the claims at hand (e.g., "we have made a definitive case"; "we have established a clear link"; "we have proved") being inducted into the canon of received knowledge even prior to their peer review and journal acceptance.

Our interest in reanalyzing Karatsolis' data archive was to raise and answer some of the questions he raised but from the vantage of our theoretical perspective and computer operationalization of that perspective. We asked: How does the language of citation differ from one discipline to the next and from one level of experience to the next? And we ask both questions from two points of view. We first ask these questions from a more "discourse-wide" (Biber, 1989; Pennebacker, 2011; Hart, 2013) perspective, where we examine how the "discourse-wide" 31 dimensions can help describe the language of citation for a particular discipline (CHEME, HSS, MSE, CS) or experience level (advisor/advisee). We then focus in on a narrower (Swales-like, 1990) "citation-specific" perspective rooted in the 13 subdimensions of appendix 2.

5. Data

As work in textual corpora goes, the Karatsolis' data set is rather small. The data set we were provided contained 27 texts written by 8 advisors and 8 advisees in four disciplines as mentioned above. In 2 of the disciplines (CHEME, MSE), the texts are balanced between advisor texts and advisee texts (3:3 and 4:4 respectively). In the remaining disciplines, the proportion of texts is 4 advisor texts: 3 advisee texts (HSS) and 4 advisee texts: 3 advisor texts (CS). There is some variation on the length of the

texts across discipline and between advisor and advisee texts within a discipline. Texts within CHEME totaled 23,612 words; within CS, 20,710 words; within HSS, 25,384 words; and within MSE, 20,990 words. Within CHEME, the 3 advisor texts accounted for 53% (12,440) of the total CHEME corpus and the 3 advisee texts accounted for 47% (11,172). Within, CS, the 3 advisor texts accounted for 37% (7,674) of the words and the 4 CS advisee texts accounted for 63% (13,036). Within HSS, the 4 advisor texts accounted for 52% (13,191) of the words and 3 advisee texts 48% (12,193). Within MSE, the two advisors accounted for 47% (9,911) of the words and the two advisees 53% (11,079).

6. Methods

6.1 Procedures

Breaking the Data Set into Paragraphs

To increase the sample size, the data were split into 734 paragraph chunks, 402 belonging to advisees and 332 belonging to advisors (Table 1).

Table 1. The breakdown of paragraphs in the data analyzed

Discipline	# Advisor Paragraphs	#Advisee Paragraphs	% Ratio: Advisor/Advisee Paragraphs
CHEME	103	101	50/50
CS	52	132	28/72
HSS	103	91	53/47
MSE	74	78	48/52

Because of the imbalance of advisors and advisees in the CS paragraphs, our research design did not include an investigation of the potential interaction effects between discipline and experience level. We instead broke down the data into two data sets for analysis looking at main effects only.

Breaking the Paragraphs into Two Subsets for Analysis

Discourse-Wide Data Set

This data set includes all the paragraphs coded from all 31 discourse dimensions, including the dimension of citation (appendix 1, dimension 2) taken as a holistic category.

Citation-Specific Data Set

This data set includes all the paragraphs coded only from 13 subdimensions of dimension 2. We make the same comparisons between discipline, role, and role within discipline as above, but now with a focus only on subdimensions specific to citation.

6.2 Statistics

Statistical Models

We applied statistical models to understand the rhetorical variation in paragraphs across both data sets by discipline (CHEME, HSS, MSE, CS), by role (advisor vs. advisee), and by role within discipline (CHEME advisor vs. CHEME advisee, and so on). We worked with the discourse-wide dataset on the original scale. To better satisfy the normality assumption of the statistical model, we worked with the citation-specific subdimension dataset on a scale transformed by the square-root function. We also dropped one outlier paragraph for both the dimension and subdimension analysis (CHEME advisee 1-IP-22) because it was the only paragraph in the entire corpus to register an instance of contingent citation ("[we] may speculate"). For ease of interpretation, we reduced the dimensions of the two datasets by fitting factor models, treating the models as interpretable "rhetorical strategies." The effect of discipline and role on these factor-strategies was then studied simultaneously by multivariate analysis of variance (MANOVA). When MANOVA results suggested that at least one pairwise comparison by discipline or role was statistically significant, we further performed multiple comparison tests on each factor-strategy to locate the source of the difference. To study the effect of role within each discipline, we examined sub-datasets within the same disciplines and studied the effect of role following similar procedures. All the comparisons were adjusted by a combination of Tukey and Bonferroni corrections to control for Type 1 errors.

Extraction and Interpretation of Three Discourse-Wide Rhetorical Strategies

Running factor analysis under conventional assumptions of factor extraction came to yield three discourse-wide factors (aka rhetorical strategies) for extraction and interpretation. We interpreted these factors as follows. Factor 1 exploited an under-observed polarity in the research article. On the one hand, research articles must show their significance by arguing for the positive and strategic contribution they make to knowledge. On the other hand, research significance stands on the foundation of research validity. And to establish the validity of their findings, research articles must typically engage in highly specialized (i.e., low-frequency) academic vocabularies that can eclipse larger significance claims. Factor 1 appeared to pick up this rhetorical polarity and extracted variables at one pole (positive values, positive relations, strategic, forceful, future) essential to establishing research significance and extracted variables on the opposite pole (specialized academic terms, exposition, description) essential to

establishing research validity. For example, consider the following paragraph from a CS advisee which scored high on the research significance side of the factor.

> By having the actual collection responsibility centralized, we provide an easy method for collaborative coordination of results. (CS Advisee 1-2-30)

The constructions "by having the" and "easy method" are recognized by the DocuScope dictionaries as strategic expressions while "collaborative" signifies positive relations and "coordination" signifies a positive value. Because of the signaling, readers outside the discipline without an accurate mental model to parse the deep semantics of the sentence can nonetheless perceive its function as a statement of claimed significance. By contrast, consider snippets of a paragraph that scored high on the research validity side of the factor.

> The alumina nanoparticles were coated similarly to the procedure found in Ref [8]. Twenty grams of nanoparticles were suspended in ethanol through 10 min of sonication (VCX-400 Sonics Materials Vibra _ cell) at 70% power. (MAT Advisee 1-1-4)

This paragraph contains specialized academic terms (e.g. alumina, nanoparticles, sonication, sonics materials), descriptive terms (e.g., ethanol, distilled water, mixture) and expository numerical expressions (e.g., 10 min., 70%). The paragraph delineates the small detail required to make the research valid, but the procedures outlined, typical of "methods" sections in technical research reports generally, mask most traces of research significance.

Factor 2 exploited another latent distinction in the research article between the foreground and background. Articles reporting original research report a project that has never before been reported and it does so from a historical record of projects that came before. Reporting prior projects form part of the "background" of the article and are populated by people, places, stories, and citations to the aforementioned. These various rhetorical devices "set the table" for the original research and provide background references to it at any point throughout the research paper. Factor 2 appeared to capture a polarity between articles that devote larger ratios to background information versus articles devoting smaller ratios. The various rhetorical devices for conveying "background" within a paragraph include past time ("was developed"), persons ("Balke and Hamielec"), places ("Greece"), public language ("is discussed") and citations ("[9, 10]"). Paragraphs from engineering with continuous citation through authorial names, bracketed numbers, past reference, and people (proper names) scored high on this factor:

> The composite polymerization procedure was developed based on the work of Balke and Hamielec [9] and is discussed in detail in a previous paper [4]. The resulting nanocomposite was compression molded into flat bars (1-mm thick) in a hydraulic press (Carver 12 ton) at 180 C and 25 mtons. (MSE Advisee 1-1-5).

As did paragraphs from the humanities with similar features combined with reference to places like Greece and North and West Britain:

> For Aristotle and Isocrates, rhetoric formed participants in the Greek polis. For Hugh Blair and Adam Smith, rhetoric formed participants in the provincial cities and towns of North Britain. For John Witherspoon and John Quincy Adams, rhetoric formed participants in a new democratic republic struggling to become something other than West Britain.

By contrast, articles scoring low on this factor featured only the dimension of "reasoning" (appendix 1 row 25), which in this context is used to advance the immediate contribution of the article shorn of background framing, which produces larger ratios of foreground over background. Consider the following paragraph from CHEME advisor 1-1P¬29 that scored high on this "high foreground/low background" pole of the factor.

Reasoning expressions in this paragraph such as the causal attribution "due to" and the cohesive marker "such" move the logic of the ideas along linearly without background framing.

> Due to the optimization framework, constraints can be explicitly imposed on both the controlled and manipulated variables...Such infeasibilities are usually handled by (1) using an infinite prediction horizon and removing the constraints in the initial portion of the prediction.

Factor 3 featured negative dimensions, including negative emotions [e.g., distress, misery; appendix 1, row 8) and negative values (e.g., injustice, treason; appendix 1, row 30) along with the dimensions of reasoning and linguistic complexity. Paragraphs scoring high on this factor coincided with a slot that Swales (1990) associated with "gaps" in the field (e.g. "there is still no definitive understanding of..."). In Swales' CARS model, gaps are areas of inconclusive research where authors seek to have impact. Our factor 3 included Swales' sense of gaps but it went broader to include any pocket of negativity that can impede or constrain research progress or that can show the negative costs and ramifications of the problem that the research hopes to address. Consider a paragraph scoring high on factor 3 where these pockets of negativity simply enumerate the challenges of the status quo:

> The chronic hyperglycemia in diabetes is associated with long-term complications due to damage, dysfunction and failure of various organs, specially the eyes, kidneys, nerves, heart and blood vessels. The three main complications being retinopathy, nephropathy and neuropathy. (CHEME advisee 2-1 unpub P-31)

With words like "damage", "dysfunction," and "failure," the paragraph above implies an urgency to treat diabetic hyperglycemia. Paragraphs in HSS that scored high on factor 3 were associated with negativity of an entirely different order and magnitude:

> Less problematic to respondents than the status markers, but still considered by Hairston to be very serious, were sentence boundary problems such as run-on sentences and

> sentence fragments, or other errors, such as not capitalizing proper names, a lack of noun-verb agreement, lack of parallelism, and faulty adverb forms. (HSS advisee 1-1 unpub P-10)

Extraction and Interpretation of Two Citation-Specific Strategies

Running factor analysis under conventional assumptions of factor extraction and interpretation recommended two citation specific factors for extraction and we interpreted these factors as follows.

Factor 1 found the conventional difference between the author-date and the numeric systems of citation. One pole loaded high on author-date citation and the other on numerical citation.

Factor 2 found a potentially more interesting mix of citation variables that is rarely discussed in the literature. The combination includes numerical citation along with authorizing-source citation ("found that"), and contestable-source citation ("argues that"). Put more simply, it involves the proximate co-occurrence of "objective" and "subjective" citation, the citation of some things as accepted fact and the citation of other things perceived to dwell in a world of argument, interpretation, and point of view. Several authors (Hyland 1999; Harwood 2009) have found that contested citation is more common in HSS disciplines than the natural sciences and engineering. But much less studied is the close intertwining of objective and subjective systems of citation. We called this factor "subjective-objective citation juxtaposition" for short.

7. Results

7.1 Results from the Discipline-Wide Data Set

In this section and the sections to follow, we examine the results from the discipline-wide data set.

Disciplinary Differences in the 3 Discourse-Wide Rhetorical Strategies

We applied MANOVA on the three discourse-wide strategies with discipline as the sole factor. The MANOVA including all three strategies showed a strong main effect for discipline, with F (9, 1769) = 13.74, $p < .001$, indicating at least one significant difference between one pairwise set of disciplines on at least one of the discourse-wide factors. The multiple comparisons test was then run on each factor to determine the source of the significance, each at an adjusted significance level of p = .05/3 or .016.

For factor 1 (research significance vs. research validity), the multi-comparisons test found differences between means that were statistically-significant in every pairwise comparison per discipline. HSS had the highest positive mean (.79) for establishing research significance and statistically higher than the second highest discipline, CS (.29). CS was statistically higher than CHEME (-.38) for establishing research significance and CHEME was statistically higher than MSE (-.84). The means for CHEME and MSE were negative, indicating that CHEME and MSE paragraphs featured more

patterns emphasizing research validity than research significance. Compared to writers of the HSS paragraphs overall, writers in the technical disciplines apparently were more likely to assume that readers could infer significance from their prior knowledge, or, alternatively, assume that readers were more interested in method over significance. The operability of either assumption would permit writers in these technical fields to assign more of their efforts to the methodological underpinnings that gave the findings validity.

For factor 2 (background vs. foreground focus), the multiple comparisons showed a statistically significant difference between the means of HSS (.39), MSE (.05), CHEME (-.11) and CS. (-.33). HSS paragraphs used textual background statistically more frequently than MSE, CHEME, or CS. MSE paragraphs used textual background statistically more frequently than CHEME or CS.

For factor 3 (negativity), multiple comparisons showed a statistically significant difference between the mean for CS (.32) and the remaining disciplines [HSS (.04), MSE (-.04) CHEME (-.22)]. In this case, CS paragraphs emphasized some combination of research gaps, negative constraints, and negative costs of the status quo more frequently than the other disciplines. These results are summarized in Table 2.

Table 2. Summary of Results for Discourse-Wide Factors vs. Discipline

Factor	Interpretation	Result
F1	Research significance vs. validity	HSS most focused on significance. HSS and CS more focused overall on significance while MSE and CHEME are more focused overall on validity.
F2	Background vs. foreground focus	HSS most focused on background. HSS and MSE are more focused on background while CS and MSE are more focused on foreground.
F3	Negativity	CS most focused on negativity. CS and HSS most focused overall on negativity while MSE and CHEME are overall less negative.

Role Differences for the 3 Discourse-Wide Rhetorical Strategies

We then applied a MANOVA on the three overall rhetorical strategies with role as the sole factor. The MANOVA including all three strategies showed no strong main effect for role, with F (3, 729) = 2.38, p = 0.069.

Role within Discipline for the 3 Discourse-Wide Rhetorical Strategies

The analysis of role within discipline required partitioning the dataset into subsets for each of the four disciplines. Accordingly, we adjusted the significance level of the Wilks' test to .05/4 or $p = .0125$. Because we had already found no significant effect for role between advisors and advisees taken as a group (7.12 above), our interest in analyzing role within discipline was to understand whether role had a significant effect within a discipline considered independently from its effect in the other three disciplines. For this reason, no further Bonferroni and Tukey adjustments on significance levels were necessary.

Advisor/Advisee Differences within CS

MANOVA was run to determine if the three discourse-wide rhetorical strategies varied by role (advisor/advisee) within CS. The MANOVA including all three factors showed no main effect between CS advisors and CS advisees in the way they traded off between significance and validity, background and foreground, and in their use of negativity, $F(3, 180) = 1.760$, $p = .157$.

Advisor /Advisee Differences within HSS

The MANOVA indicated a strong main effect for role, $F(3, 190) = 9.907$, $p < .001$ within HSS. Multiple comparisons showed that HSS advisees had significantly higher means on negativity (factor 3) than HSS advisors. Table 3 lists the mean scores on all three factors for both HSS advisees and HSS advisors.

Table 3. Means for HSS Advisees/Advisors on 3 Discourse-Wide Factors

Factor	Interpretation	Advisee	Advisor
F1	Research significance vs. validity	.93	.66
F2	Background vs. foreground focus	-.33	-.31
F3	Negativity*	.41	.07

* = significant difference ($p < .01$)

Advisor/Advisee Differences within MSE

The MANOVA including all three rhetorical strategies showed no main effect for role in MSE, $F(3, 148) = 1.521$, $p = .211$.

Advisor /Advisee Differences within CHEME

The MANOVA including all three rhetorical strategies showed a main effect for role [$F(3, 199) = 5.691$, $p < 0.001$] and multiple comparisons showed a main effect on factor

3 (negativity) with CHEME advisors having means for negativity (.03) significantly higher than the means of CHEME advisees (-.45).

Table 4. Summary of Results for Discourse-Wide Factors vs. Role

Factor	Interpretation	Result
F1	Research significance vs. validity	**No main effects** between advisors/advisees within HSS, CS, MSE, CHEME
F2	Background vs. foreground focus	**No main effects** between advisors/advisees within HSS, CS, MSE, CHEME
F3	Negativity	**Main effects** between advisors and advisees **only in HSS and CHEME**. In HSS, advisees are significantly more negative than advisors. In CHEME, advisors are significantly more negative than advisees.

7.2 Results from the Citation-Specific Subdimension Data Set

In this section and the sections to follow, we examine the results from the citation specific data set. We will soon delve deeper into the analyzing the two factors extracted from the citation-specific data set (see Section 6.24), but before doing that, we focus on 3 citation-specific variables (Appendix 2) that overlap substantially with citation variables studied in the literature.

Preliminaries: Using Three Citation-Specific Variables to Calibrate our Citation-Specific Measures with Previous Literature

To calibrate whether these three citation-specific measures corresponded with measures already used in the literature, we sought to see if we could replicate previous findings. The variables in question are countering sources, contestable sources, and authorizing precedent (see Appendix 2 for definitions). Humanities disciplines have been characterized as more "disputational" (Hyland, 1999, 362) than the sciences and with a "slower" growth-rate and a greater propensity for historical (precedent) citation than engineering (Halevi, 2013). Accordingly, we would predict that HSS paragraphs reflect these differences by showing more countering, contestable, and authorizing precedent citation than the citations of the non-HSS disciplines. The MANOVA text including these three specific variables showed a significant main effect for discipline, F (9, 1769) = 6.635, $p < .001$. Multiple comparisons confirmed the direction of these findings as reported in the literature: HSS paragraphs had the highest means for all three citation types. For contestable sources, HSS writers showed means (.09) significantly higher than MAT (.013), CS (.011) and CHEME (.008); for countering citation, the difference in means between HSS (.045) and MSE (.004) and CHEME (.000) were significant. However, the difference between HSS and CS (.015) was not significant.

For authorizing precedent citations, HSS writers (.046) were significantly higher than CHEME writers (.000) and higher than MSE (.030) and CS writers (.026) but not significantly so. Table 6 illustrates the dominance of these three citation-specific variables in HSS writing.

Table 5. Mean Differences on 3 Citation-Specific Subdimensions /Discipline

Means	HSS	MSE	CHEME	CS
Authorizing Precedent	.05	.03	0.0	.03
Contestable Sources	.09	.01	.01	.01
Countering Sources	.05	.00	.00	.02

Although these three citation-specific variables strongly distinguish HSS from the other disciplines, they had no effect distinguishing advisors and advisees across the four disciplines [F (3, 729) = .263, p = .852].

Disciplinary Differences for the 2 Factorized Subdimensions

MANOVA was run to determine if the two citation-specific factors (section 6.24) varied by discipline. The MANOVA including both factors showed a strong main effect for discipline [F (6, 1456) = 23.602, p < .001]), indicating at least one significant difference on one of the citation-specific factors between one pairwise set of disciplines.

For factor 1 (author/date vs. numerical citation reference), the more positive the average the more a discipline favored author-date citation. The more negative the average, the more a discipline favored numerical citation. The multiple comparisons test showed significantly different means between HSS paragraphs (HSS mean = .478) and the means of CHEME (CHEME mean = .07), CS (CS mean = -.23) and MSE (MSE mean = -.42). Unsurprisingly, HSS paragraphs relied on author-date citation more than the technical disciplines. But the technical disciplines did use author-date citation and HSS writers did use numerical citation. Curiously, the paragraph scoring highest on [single] numeric citation came from an HSS writer writing on composition.

> This modeling approach comes very close to a process described by Phelps [14] who articulated an approach to structural analysis drawing on and responding to work by Faigley & Witte [15] and Van de Kopple [16] in composition studies, as well as Halliday & Hassan [17] and Van Dijk [18] in linguistics. (HSS advisor 1-1P-25)

But the placement of parenthetical citation in this paragraph is less curious when we notice it relies on integral citation (author information is part of the sentence syntax) rather than non-integral (author information stands outside the sentence syntax; see

Swales 1990). In his study of citations across disciplines, Hyland (1999) found that technical fields rely on non-integral citations in order to emphasize the research more than the agents conducting it. Hyland (1999) and Mansourizadeh and Ahmad (2011) found that integral-citations, as in the HSS paragraph above, are more popular in the so-called "soft disciplines" where writers seek more opportunity to express "stance" and "make evaluations" (Mansourizadeh and Ahmad, 2011, 153).

For factor 2 (objective/subjective citation juxtapositions), multiple comparisons showed a statistically significant difference between the factor score means of CHEME (.27) and CS. (.20) on the one hand and the means of HSS (-.22) and MSE (-.32) on the other. In this case, the more negative the mean score, the more likely the discipline contained paragraphs that included both objective (authorizing) and subjective (contestable) citations. This juxtaposition of objective and subjective citation happened in paragraphs, such as HSS advisee 2-1, unpub P-13, who in the same paragraph cites the authority of collected "interviews," "conversations," and "written materials" along with what is "alleged" and "claimed" about these materials that cannot be taken on their face. The juxtaposition appears in MSE advisor 2-IP-0, which contains authorizing citation verbs like "have been used" and "showed" with contestable citation verbs like "it has been argued." It further appears in paragraphs, such as CHEME advisee 2-1 unpub P-27, which contained both the authorizing "X modeled" and the contestable citation "X proposed."

Table 6. Summary of Results for Citation-Specific Factors vs. Discipline

Factor	Interpretation	Result
F1	Numerical vs. Author-Date Citation	**Main Effect:** HSS writers used Author-Date Citation significantly more than writers in the other disciplines.
F2	Juxtaposition of Objective/Subjective Info	**Main Effect:** Writers in HSS and MSE used subjective/objective juxtapositions significantly more than writers in CHEME and CS.

Role Differences for the 2 Factorized Citation-Specific Subdimensions

MANOVA was run to determine if the two citation-specific factors varied by role (advisor/advisee). The MANOVA including both factors showed a strong main effect for role [F (2, 730) = 13.328, $p < .001$]. Multiple comparisons showed that the paragraphs contained significantly higher means for advisees (.16) than for advisors (-.20) on factor 1, where a negative score meant a higher use of numerical citation and a positive score a higher use of author-date citation. In this case, advisors were likely to use numerical citation more than advisees. There was no significant effect, however, for role when it came to combining objective and subjective citation. The mean factor scores were

small for both groups [advisor: -03; advisee .02], which suggests that combining objective and subjective citation was not a frequent phenomenon for either advisors or advisees.

Role within Discipline for 2 Factorized Citation-Specific Subdimensions

Advisor/Advisee Differences within CS

MANOVA was run to determine if the two citation specific factors varied by role (advisor/advisee) within CS. The MANOVA including both factors showed no main effect for role, F (2, 181) = .403, p = .669.

Advisor /Advisee Differences within HSS

MANOVA indicated a strong main effect for role [F (2, 191 = 9.44, p = .000)]. Multiple comparisons showed that HSS advisors had significantly higher means on factor 1 (author-date vs. numerical citations) than HSS advisees. Positive means indicated disproportionately author-date citation and negative means indicates disproportionately numerical citation. The means for HSS advisors were .66 and for HSS advisees .93. HSS advisors mainly contributed published articles to Karatsolis' corpus and advisees mainly contributed manuscripts. It may be that numerical citation is more reflective of published formats, which would mean that the effect for role interacts with an effect for genre and format (published vs. unpublished manuscript). Multiple comparison tests showed no significant difference between advisors and advisees on factor 2, the juxtaposition of objective and subjective citation.

Advisor/Advisee Differences within MSE

MANOVA including both factors showed no main effect for role within MSE [F (2, 149) = 2.096, p = .127.

Advisor /Advisee Differences within CHEME

MANOVA was run to determine if the two citation-specific factors varied by role (advisor/advisee) within CHEME. The MANOVA including both factors showed no main effect for role [F (2, 201) = 2.174, p = .116].

Table 7. Summary of Results for Citation-Specific Factors vs. Role

Factor	Interpretation	Result
F1	Numerical vs. Author-Date Citation	**Main Effect:** HSS advisors used numerical citation significantly more than HSS advisees.
F2	Juxtaposition of Objective/Subjective Info	No Main Effect

8. Discussion

8.1 Findings and Significance

Our findings show the promise of using dictionary methods to study citation patterns. Part of this promise lies in the capacity of dictionary methods to independently confirm or complement manual coding studies. While our findings uncovered many regularities not found in Karatsolis' hand codings, we took our results as complementary to Karatsolis' findings and not at all at odds with them. In addition, using dictionary methods, we were able to independently replicate some findings of previous literature, namely the greater frequency of contested, countering, and historical citation in HSS writing.

But an even greater promise lies in the ability of automatic methods to discover fresh and quantifiable textual constructs too diffusely distributed across a text to be identified (or counted) through serial reading and so unlikely to find their way into a manual coding scheme. Examples of these distributed patterns are research significance vs. research validity, the textual foreground vs. background, and degrees of negativity. Based on our common data source, these constructs significantly contribute to capturing important differences in academic writing and citation by discipline. We found that HSS disciplines are particularly distinct over engineering in the way they require writers to expand on significance and background. More than the technical disciplines, HSS writers have to make a case for the legitimacy of a problem before they embark on efforts to address it.

As we saw with the DocuScope dimensions that constituted significance (positive values, positive relations, strategy, forceful, future) and backgrounding (people, places, stories), an academic text's significance and background can't be achieved without tying ideas to persons, places, historical systems, communities, strategies, and values. The more technical disciplines may negotiate a more blurred line between research significance and technical virtuosity, meaning that the apparatus that triggers the decision to cite in HSS fields may be less salient citation triggers in technical fields. This does not reduce the need for citation in technical fields. But it may narrow the decision-to-cite to prior work already focused on like-minded virtuosity on similarly-related problems.

We also found that computer science dominated the other disciplines in negativity – gaps, constraints, and costs – which can be inconclusively interpreted as rhetorical gestures to attract importance and funding. The juxtaposition of objective (fact-based) and subjective (opinion-based) is another pattern not likely to be discovered through serial reading. It was not the basis of main effects in Karatsolis' data. This is probably due in part to the strong "objective" flavor of the texts he collected. In the common data we analyzed, subjective citation (median = 0) was the weak partner of objective citation (median = .71). One can imagine more "contentious" corpora where subjective citation equaled or even trumped objective citation. While it was beyond the scope of this paper to analyze all the possible meanings of this juxtaposition in Karatsolis' data,

there is reason to believe that this co-occurrence of objective and subjective citation types may prove an important explanatory variable in larger corpora.

8.2 Dictionary Methods: Acknowledging Real Limits and Debunking Prejudices

Automatic dictionary methods have limitations and their greatest limitation is also their strength if the limitations are properly contextualized: automatic dictionaries are incapable of reading a text or providing an adequate replacement for the serial reader. The strength of automatic dictionaries is to find patterns inaccessible to the serial reader. But the ultimate value of what an automatic dictionary finds depends on the close inspection of the serial reader to "read beneath" the patterns discovered and seek to authenticate the results through the triangulation of independent research and scholarship. Using automated dictionaries without scholarly backup can lead to unsupported and potentially disastrous inferences. Automatic dictionaries point out new distributed patterns in a text for researchers to attend to, but researchers must give these patterns their deep serial reading and scholarly attention to know if statistical aggregation of these patterns actually advances knowledge. The best practitioners in dictionary approaches to language study (Hart, Pennebaker) proceed with these cautions.

Overall, based on our measures, we found that discipline had a consistently greater impact than role in defining the rhetorical behaviors of citation in Karatsolis' data. Even when we did find significant pattern differences between advisors and advisees across or within disciplines, the pattern differences amounted mainly to the system of citation used. But this limitation in our finding for role likely speaks to the deepest limitations of a pure dictionary approach to the study of citation. As Karatsolis (2005) found, as well as researchers who came before (Dong, 1996; Thompson and Tribble, 2001) and after him (Petrié, 2007; Harwood, 2009; Mansourizadeh and Ahmad, 2011), quantitative patterns that count citations and categorize them into rough-grained categories lack the precision to capture what might be called citation "acumen," the deep strategies separating novices and experts that Karatsolis (2005) and Harwood (2009) only began to uncover when they moved from simple citation reference counts to more qualitatively fine-grained interpretative coding schemes and discourse-based interviews. Since role (advisor/advisee) differences are essentially novice/expert differences, dictionary approaches to the study of citation that rely only on surface texts are likely to be limited. As Karatsolis and Harwood ended up doing, we felt we would need more fine-grained qualitative approaches and access to the writers themselves (not just their texts) to dig deeper in the question of citation acumen as a developmental strategy.

This limitation of automated dictionary methods constitutes a limitation from within. But automated dictionary methods as methods for writing research also face external criticism from certain proponents of machine learning methods. Many of the criticisms of this community are well-founded. Building reliable and robust dictionaries

is labor-intensive and error-prone work. But perhaps the most severe criticism is that dictionaries, built in one context, are impervious to classifying accurately in new contexts. They can't learn and adapt to new contexts and, for this reason, succumb to the very limitations that machine learning methods were specifically designed to overcome (Grimmer and Stewart, 2013). As Loughran and McDonald (2011) note, words can shift their valence between positive and negative depending on context and ideology. "Tax" can be a bad word in the phrase "tax and spend" but a contributor to the positive phrase "tax relief." In polite company, "crude" is a word of opprobrium; among oil executives, it is a word of profit. Politicians during the cold war spoke ominously of the "cancer of communism." That sense of boding may persist when an oncologist tells a patient she has cervical cancer. But when that same oncologist reads about cervical cancer in a medical book, the context shifts and "cancer" becomes a dispassionate object of study rather than an ominous reference.

Critics of dictionary methods often cite the contextual serendipity of words as a fatal blow against human-crafted and so-called "fixed" dictionary approaches. As we have tried to show here, managing contextual serendipity is completely compatible with dictionary approaches so long as the dictionary includes a flexible development environment with flexible tools for letting dictionaries adapt, evolve, and expand in complexity in order to handle an increasingly large array of contexts. No one would complain if machine learning approaches could truly handle contextual serendipity as effortlessly as the human brain. But we are some way off before machine learning will grow to that level of sophistication. At the same time, there is no reason to dichotomize dictionary and machine learning approaches. We believe they capture complementary approaches that can in principle converge and collaborate. Humans process ideas from text serially and deeply but can't keep track of hundreds of variables at once. Machines can process many variables at once, but without local precision and historical depth. Currently, supervised approaches in machine learning require the acumen of human coders to learn and generalize from. The melding of dictionary and machine learning approaches simply makes human annotation a more visible and more equal partner in the effort to unleash machines on texts for meaning.

9. Conclusion

This paper has presented a dictionary approach to text analysis and an illustration of the benefits of applying new approaches to existing and previously analyzed data sets. Just as importantly, it has given us an opportunity to participate in a project with research teams relying on different theoretical frameworks and methods to analyze a common set of texts. Our participation has left us with important positive lessons. It has helped us make sense of our findings and the strengths and limits of our own approach. As this subsection shows, approaches to writing research are governed by the tradeoffs of richness and scale. Dictionary approaches scale better than manual coding but scale less well than machine-based NLP approaches. At the same time, dictionary

approaches afford a "richness" of context and construal over machine-based approaches, but prove less rich than manual methods. As Geisler notes in her closing section, interpretation plays a role across all the approaches, but in different ratios and along different schedules. We believe each approach benefits from a self-awareness of its positioning relative to other approaches and provides openings for mutual acceptance and even integration. In sum, conducting our research on citation with data shared from other researchers taking different perspectives convinced us of the deep benefits of using a common dataset for writing research. Common data analysis, we believe, opens a path for writing researchers to advance the field more coherently because we are advancing it from common baselines of attention.

Acknowledgements

The authors wish to thank Cheryl Geisler, Luuk Van Waes, and the anonymous reviewers for their editorial attention and their useful feedback that improved the quality of this paper. All remaining errors are the responsibility of the authors. This research was undertaken under the umbrella of the Simon Initiative for Technology Enhanced Learning in Writing/Communication at CMU, http://www.cmu.edu/simon.

Notes

1. https://owl.english.purdue.edu/
2. http://wpacouncil.org/positions/WPAplagiarism.pdf
3. In 1993, Rebecca Moore Howard first coined the term "patch writing" to mean "copying from a source text and then deleting some words, altering grammatical structures, or plugging in one-for-one synonym-substitutes." It means in effect copying/converting isolated sentences rather than copying/converting deeper meanings from the sources that are conveyed through and across sentences. For years, Moore believed that patch writing resulted from students' genuine efforts to struggle with difficult material expressed in long, difficult sentences. It was only when she teamed up with Sarah Jamieson and started coding student citation practices did she learn that her initial views were incomplete. Coding citations in 174 research papers from students across 16 institutions, Jamieson found that students often patch wrote from "short, simple" sentences and that over 69% of students concentrated all their citations in the first two pages of the source text. Such findings forced Howard and Jamieson to see student reading comprehension and engagement as contributors to student patch writing and other faulty citation practices. See Howard, 2014 and Jamieson, 2013.
4. To be fair, some of Biber's more recent work has encouraged the investigation of particular texts. See Biber and Conrad, 2009.
5. DocuScope produces four types of data--counts, frequencies, ratios, and characters. Counts are self-explanatory. Frequencies normalize for length by dividing the number of pattern hits by the total number of words per text. Ratios consider the % of a particular DocuScope category used as a ratio of all other categories used. Thus, if negative emotion in a text is .05% of all the categories used, the ratio score of negative emotion is .05%. Ratio score was inspired by Kenneth Burke's insight that we can gain much insight about how a text functions rhetorically by considering the ratio of a particular measure of interest against the whole set of possible measures (Burke 1969: 228). Characters gives weight to every character in a text and

so gives additional rhetorical weight to longer words over shorter words. In this paper, all measures were frequencies unless we specifically refer to them as counts.

6. http://www.flintbox.com/public/project/23053

References

Al-Malki, A., D. Kaufer, et al. (2012). *Arab Women in Arab News*. London, Bloomsbury.

Bawarshi, A. (2003). *Genre, & the Invention of the Writer*. Logan, Utah State University.

Bazerman, C. (1987). Codifying the social scientific style: The APA Publication Manual as a behaviorist rhetoric. In S. Nelson, A. Megill, & D.N. McCloskey, *The rhetoric of the human sciences: Language, & argument in scholarship, & public affairs* (pp.125-144). University of Wisconsin.

Biber, D. (1989). *Variation in Speech, & Writing*. New York, Cambridge University Press.

Biber, D., S. Conrad, et al. (2004). If you look at...Lexical Bundles in University Teaching, & Textbooks. *Applied Linguistics 25*(3), 371-405. doi: 10.1093/applin/25.3.371

Biber, D., & S. Conrad (2009). *Register, Genre,, & Style*. New York, Cambridge. doi: 10.1017/cbo9780511814358

Burke, K. (1969). *A Grammar of Motives*. Berkeley, University of California Press.

Bybee, J. (2007). *Frequency of Use, & the Organization of Language*. New York, Oxford University Press. doi: 10.1093/acprof:oso/9780195301571.001.0001

Collins, J., D. Kaufer, et al. (2004). Detecting Collaborations in Text. Comparing the Authors' Rhetorical Language Choices in the Federalist Papers. *Computers in the Humanities 15*(1), 15-36. doi: 10.1023/B:CHUM.0000009291.06947.52

Cotos, E. (2014). *Genre-based Automatic Writing Evaluation for L2 Research*. New York: Palgrave Macmillan.

Cotos, E., Huffman, S. & Link S. (2015). Furthering, & applying move/step constructs: Technology-driven marshalling of Swalesian genre theory for EAP pedagogy. *Journal of English for Academic Purposes. 19,* 52-72. doi: 10.1016/j.jeap.2015.05.004

Cronin, B. (1984). *The Citation Process*. London, Taylor Graham.

Council on Writing Program Administrators. *Defining, & Avoiding Plagiarism: The WPA Statement on Best Practices*. Retrieved from http://wpacouncil.org/positions/WPA plagiarism.pdf.

Devitt, A. (2008). *Writing Genres*. Carbondale, Southern Illinois.

Dong, Y. R. (1996). Learning How to Use Citations for Knowledge Transformation: Non Native Doctoral Dissertation Writing in Science. *Research in the Teaching of English 30*(4), 428-457.

Geisler, C. (1994). *Essayist Literacy, & the Nature of Essayist*. New York, Routledge.

Gilbert, G. N. (1977). Referencing as persuasion. *Social Studies of Science 7,* 113-122 doi: 10.1177/030631277700700112

Gilbert, G. N. (1976). The Transformation of research findings into scientific knowledge. *Social Studies of Science 6,* 281-306. doi: 10.1177/030631277600600302

Gilbert, G. N., & M. Mulkay (1984). *Opening Pandora's Box: A Sociological Analysis of Scientist's Discourse*. Cambridge, Cambridge University Press.

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago, University of Chicago.

Grimmer, J., & B. M. Stewart (2013). Text as data: The promises, & pitfalls of automatic content analysis methods for political texts. *Political Analysis 21*(3), 267-297. doi: 10.1093/pan/mps028

Halevi, G. (2013) *Citation characteristics in the Arts &* Humanities. Research *Trends 32*. Retrieved August 4, 2015 at http://www.researchtrends.com/issue-32-march-2013/citation-characteristics-in-the-arts-humanities-2/.

Hart, R. P. (2015). Genre, & Automated Text Analysis: A Demonstration. In J. Ridolfo, & W. Hart-Davidson, *Rhetoric, & the Digital Humanities* (pp. 152-168). Chicago, University of Chicago Press: 152-168.

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

Hayes, J. R., & D. Baizek (2008). Understanding, & Reducing the Knowledge Effect. Written Communication 25(1): 104-118. doi: 10.1177/0741088307311209

Hope, J., & M. Witmore (2007). Shakespeare by the Numbers: On the Linguistic Texture of the Late Plays. In S. Mukherji, & R. Lyne, *Early Modern Tragicomedy* (pp. 133-153). London, Boydell, & Brewer.

Hopper, P., & E. Traugott (2003). *Grammaticalization* (2nd Edition). Cambridge, Cambridge University Press. doi: 10.1017/CBO9781139165525

Howard, R. M. (1993). A plagiarism pen-timento. *Journal of Teaching Writing, 11*, 233–245.

Howard, R. M. (2014). Why This Humanist Codes. *Research in the Teaching of English, 49I*(1), 75-81.

Howard, R., & A. Robillard (2008). *Pluralizing Plagiarism: Identities, Contexts, Pedagogies*. Portsmouth NH, Heinemann.

Hyland, K. (1998). Persuasion, & Context: The Pragmatics of Academic Metadiscourse. *Journal of Pragmatics 30,* 437-455. doi: 10.1016/S0378-2166(98)00009-5

Hyland, K. (1999). Academic Attribution: Citation, & the Construction of Disciplinary Knowledge. *Applied Linguistics 20*(3): 341-367. doi: 10.1093/applin/20.3.341

Hyland, K. (2004). *Disciplinary Discourses: Social interactions in academic writing*. Ann Arbor, University of Michigan Press.

Hyland, K. (2005). *Metadiscourse: Exploring Interaction in Writing*. New York, Continuum Press.

Hyland, K (2012). *Disciplinary Identities: Individuality, & Community in Academic Discourse*. Cambridge: Cambridge University Press.

Iowa State, Writing Center. (2015). *Citation Education*. Retrieved at April 23, 2015, from http://www.dso.iastate.edu/wmc/students/citationed.

Ishizaki, S., & D. Kaufer (2011). Computer-Aided Rhetorical Analysis. In P. McCarthy, & C. Boonthum, *Applied Natural Language Processing, & Content Analysis: Identification, Investigation,, & Resolution* (p. 276-296). Hershey PA, IGI Global.

Jamieson, S. (2013). Reading, & engag¬ing sources: What students' use of sources reveals about advanced reading skills. *Across the Disciplines, 10*(4). Retrieved from http://wac.colostate.edu/atd/reading/jamieson.cfm.

Jamieson, S. & Howard, R.M. (2013). Sentence-mining: Uncovering the amount of reading, & reading comprehension in college writers' researched writing. In Randall McClure & James P. Purdy (Eds.). *The new digital scholar: Exploring, & enriching the research, & writing practices of next gen students* (pp. 111-133). Medford, NJ: Information Today.

Karatsolis, A. (2005). *Synthesizing from Sources: Patterns of Citation Use in the Academia, & Implications for the Design of Electronic Reading, & Writing Systems. Language Literature Communication*. Troy NY, Rensselaer Polytechnic Institute. PHD Dissertation.

Kaufer, D., & C. Geisler (1989). Novelty in Academic Writing. *Written Communication 6*(3): 286-311. doi: 10.1177/0741088389006003003

Kaufer, D., & K. Carley (1993). *Communication at a Distance: Exploring the Influence of Print on Sociocultural Interaction, & Change*. New York, Routledge.

Kaufer, D., & R. Hariman (2008). A Corpus Analysis Evaluating Hariman's Theory of Political Style. *Text & Talk 28*(4): 475-500. doi: 10.1515/TEXT.2008.023

Kaufer, D., S. Ishizaki, et al. (2004). *The Power of Words: Unveiling the Speaker, & Writer's Hidden Craft*. New York:, Routledge.

Kaufer, D., S. Ishizaki, et al. (2004). Teaching Language Awareness in Rhetorical Choice Using IText, & Visualization in Classroom Genre Assignments. Jou*rnal for Business, & Technical Communication 18*(3): 361-402. doi: 10.1177/1050651904263980

Kaufer, D. (2006). Genre variation, & minority ethnic identity: Exploring the Personal Profile in Indian American Community Publications. *Discourse & Society, 17*(6): 761-784. doi: 10.1177/0957926506068432

Kennedy, K., & Howard, R. M. (2013). Introduction to the Special Issue on Western Cultures of Intellectual Property. *College English,* 75(5), 461-469.

Knuth, D., J. H. Morris, et al. (1977). Fast pattern matching in strings. *Journal on Computing* 6(2): 323-350. doi: 10.1137/0206024

Leijten, M., Van Waes L., Schriver, K., & Hayes, J.R. (2014). Writing in the workplace: Constructing documents using multiple digital sources. *Journal of Writing Research, 5*(3), 285-337. doi: 10.17239/jowr-2014.05.03.3

Loughran, T., & B. McDonald (2011). When is a liability not a liability? Textual analysis, dictionaries,, & 10-Ks. *Journal of Finance 66*(1): 35-65. doi: 10.1111/j.1540-6261.2010.01625.x

Lunsford, A. Fishman, J., & Liew W. (2013). College Writing, Identification,, & the Production of Intellectual Property: Voices from the Stanford Study of Writing. *College English,* 75(5), 470-492.

Mansourizadeh, K., & U. K. Ahmad (2011). Citation practices among non-native expert, & novice scientific writers. *Journal of English for Academic Purposes 10,* 52-161. doi: 10.1016/j.jeap.2011.03.004

McInnis, R., & D. Symes (1988). David Riesman, & the concept of bibliographic citation. *College, & Research Libraries 50,* 387-399. doi: 10.5860/crl_49_05_387

Mertha, A. C. (2007). The Politics of Piracy: Intellectual Property in Contemporary China. M.H. MacRoberts, & B. R. MacRoberts (1986). *Quantitative measures of communication in science: A study of the formal level. Social Studies of Science (16)* (pp. 151 172). Ithaca, Cornell University Press.

Moravscik, M. J., & P. Murugesan (1975). Some Results on the Function, & Quality of Citations. *Social Studies of Science 5,* 86-92. doi: 10.1177/030631277500500106

Pennebaker, J. W. (2011). The Secret Life of Pronouns. New York, Bloomsbury

Perelman, C., & L. Olbrects-Tyteca (1969). *The New Rhetoric: A Treatise on Argumentation*. Notre Dame, University of Notre Dame Press

Petrié, B. (2007). Rhetorical functions of citations in high-, & low-rated master's theses. *Journal of English for Academic Purposes 6,* 238-253. doi: 10.1016/j.jeap.2007.09.002

Pinker, S. (2014). *The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century*. New York, Penguin.

Price, D. (1963). *Little Science*, Big Science. New York: Columbia University Press.

PSSA (2015). *The Pennsylvania System of School Assessment: English Language Arts Preliminary Item, & Scoring Sampler*. Harrisburg, PA, Pennsylvania Department of Education Bureau of Assessment, & Accountability.

Ritter, K. (2005). The Economics of Authorship: Online Paper Mills, Student Writers,, & First-Year Composition. *College Composition, & Communication 56*(4), 601–31.

Scardamalia, M., & Bereiter, C. (1987). Knowledge telling, & knowledge transforming in written composition. In S. Rosenberg (Ed.), *Advances in Applied Psycholinguistics, Vol. 2: Reading, Writing, & Language Learning* (pp. 142-175). Cambridge: Cambridge University Press.

Small, H. G. (1978). Cited documents as concept symbols. *Social Studies of Science 8,* 327-340. doi: 10.1177/030631277800800305

Southwestern University. *Debbie Ellis Writing Center. Tips for Avoiding Accidental Plagiarism.* Retrieved from http://www.southwestern.edu/offices/writing/students/index.php.

Swales, J. W. (1990). *Genre Analysis: English in Academic, & Research Settings*. Cambridge, Cambridge University Press.

Thompson, G., & Y. Ye (1991). Evaluation of the Reporting Verbs Used in Academic Papers. *Applied Linguistics 12,* 365-382. doi: 10.1093/applin/12.4.365

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

Weinstock, M. (1971). *Citation Indicies. Encyclopedia of Library, & Information Science*: 16-40.

Wenger, E. (1998). *Communities of Practice: Learning, meaning, & identity*. Cambridge, Cambridge University Press.

Appendix 1: 31 "discourse-wide" Dimensions associated with the DocuScope Text Analysis and Visualization Environment. The left column contains the dimension name; the right column contains primary associations of the dimension followed by actual words and phrases (in italics) that constitute it.

	Dimension	Primary Associations: *Example Words/Phrases*
1	academic	low-frequency, specialized, abstract: *hegemonic, discursive*
2	citation	referencing second-hand-authority: *according to, argues that*
3	cohesion	linking by addition, similarity, contrast: *moreover, similarly*
4	comparison	The like of comparison or resemblance: *more/less than, like a*
5	contingency	conditioned, depending-on, probable: *if, contingency, befalling*
6	description	observed, sensed, tangible, concrete: *cat, dog, table*
7	directing movement	manual instructions and procedures: *grab, drag, fold, insert*
8	emotion-negative	negative-valence: distress, misery, blah
9	emotion-positive	positive-valence: happy, wonderful, exuberant
10	exposition	defining, illustrating, specifying: *defined as, for example*
11	facilitate	enable, guide, invite, request, suggest, recommend
12	first-person	self-reference, ego-involved: *I-me-my-mine*
13	forceful	command, confidence, insistent, intense, emphatic: *must come,*
14	future	yet-to-materialize: forecast, predict, project, portend
15	inquiry	curious, interesting, involving, puzzle, probing, mystery
16	interactive	addressing others, second person, questions: *you, would you*
17	linguistic complexity	(anti-narrative) embedded NPs, subordinators/coordinators
18	narrative	transitive-time, temporal adverbs, past verbs: *came-saw-conquered*
19	opposition	opposing, challenging, obstructing, resisting, withholding support
20	past	already-happened-and-done: *years ago, in the past*
21	persons	named entities: Sally, Mao, CIA, Teamsters
22	place	regions, cities, states, capitals: *city, district, municipality*
23	privy	subjective-private-personal: *confessed, disclosed*
24	public	institutional, activism, authoritative, bureaucratic: *legislative branch*
25	reasoning	premise/conclusion: thus, therefore, due to, owes to
26	relations-positive	liking, love, friendship, buddies, solidarity
27	relations-negative	social-division, dislike, enmity, rivals, envied
28	reporting	updates-events-change: announced, declared, transformed
29	strategic	advancing/blocking goals through plans: *plans, goals, shrewd*
30	values-negative	failed standards-to-renounce: *injustice, unfairness*
31	values-positive	approved standards-to-uphold: *justice, fairness*

Appendix 2: 13 "Citation-Specific" Subdimensions of Dimension 2

CITED AUTHORITY = CITING A LONGSTANDING OR TRUSTED KNOWLEDGE SOURCE

01. Authorizing Sources; e.g. ("is widely believed"; "was substantiated by").

02. Authorizing Precedent; e.g. ("is a long tradition"; "has long been judged")

CITED CLAIMS = CITING UNCONFIRMED, CONTINGENT OR CONTESTED KNOWLEDGE SOURCES

03. Contestable Sources; e.g. ("is widely debated"; "she argued for")

04. Contingent Sources; e.g. ("she may have shown"; these findings could")

05. Countering Sources; e.g. ("wrong to think"; "contradicts previous research")

06. Self-Citing Unconfirmed Current Work; ("we have established"; "we have shown")

CITED REFERENCE = SIGNALING A CITATION

07. Cited Author-Dates; e.g. Jones (1969)

08. Cited Author-Dates [Multiple]; e.g. [Jones, 1969; 1972]; e.g. [Jones 1969; Smith, 1985)

09. Cited Numerical Citation; e.g., [1]

10. Cited Numerical Citations [Multiple]; e.g. [1, 2, 5, 23]

11. Cited Pages; e.g. (pps 23-49)

CITED GAPS = CITING GAPS IN THE LITERATURE

12. Cited gaps in the literature; e.g., ("still not well understood"; "requires more research")

CITED QUOTATION = CITING DIRECT SPEECH

13. Cited Quotation; e.g. ("ask not what your country can do for you....")