Finding Genre Signals in Academic Writing

Ryan Omizo* & William Hart-Davidson°

* University of Rhode Island | USA

° Michigan State University | USA

Abstract: This article proposes novel methods for computational rhetorical analysis to analyze the use of citations in a corpus of academic texts. Guided by rhetorical genre theory, our analysis converts texts to graph-theoretic graphs in an attempt to isolate and amplify the predicted patterns of recurring moves that are associated with stable genres of academic writing. We find that our computational method shows promise for reliably detecting and classifying citation moves similar to the results achieved by qualitative researchers coding by hand as done by Karatsolis (this issue). Further, using pairwise comparisons between advisor and advisee texts, valuable applications emerge for automated computational analysis as formative feedback in a mentoring situation.

Keywords: citation, computational rhetoric, rhetorical moves, text processing

journal of WRITING RESEARCH

Omizo R.,& Hart-Davidson W. (2016). Finding genre signals in academic writing. *Journal of Writing Research*, *7*(3), 485-509. doi: 10.17239/jowr-2016.07.03.08

Contact: Ryan Omizo, University of Rhode Island, 101 Davis Hall, Kingston, RI 02881 | USA – rmomizo@uri.edu

1. The Challenges of Learning "Always Custom" Academic Genres

Pare' (2015) calls our attention to a seemingly paradoxical quality of genres of academic writing that, when we pause to consider it, accounts for the challenge many of us feel as novice writers. Certain types of texts, at a broad level, and more importantly the opportunities to produce these certain types of texts, recur. Drawing upon rhetorical genre theory (RGT), however, Pare notes that repetition and the expectation that there will be stable textual patterns that arise from repetition is disrupted by another phenomenon: the "always custom" nature of any one instance of an academic genre. Citing Bakhtin, Pare explains that while the instances in which certain texts arise and recur may be relatively stable, the specific circumstances, not to mention the creativity of individual writers, allow and sometimes demand that each text varies from any other. First paragraph text is not indented.

RGT prepares us for this "always custom" nature of genres by implying some distance between the genre - a term usually used by rhetorical genre theorists to signify the recurring social actions from which textual regularities arise - and a genre instance, or a single text that emerges from the distinctive social exigencies associated with a particular genre. But this very distance - the variability of any given instance of a genre from any other instance - poses challenges for learners and teachers, advisors and advisees. Learning to write in a new genre is more than just getting the words right. It is getting the rhetorical moves associated with the genre right. And that means the words can sometimes vary quite a lot!

Pare sums up the nature of the challenge succinctly with a conjecture: "Perhaps the most important lesson of RGT is that the repeated texts we assign or investigate are merely the centre of much larger patterns of typified action" (A-91). Following Pare, we would extend the reading of Bakhtin's conception of textual conventions as being part of, and indeed constitutive of, social relations by pointing out that where there are repetitions at a textual level, these can be understood as signals shared by author and reader about the social activity - the genre - they are co-negotiating. The repeated text may include large passages or just a handful of words. But it is not the mere repetition of the words, themselves, that is significant. If we look at two texts by the same writer, we might see her substitute many words and still be recognized as producing, in both cases, a recognizable (but custom) instance of the genre.

Given this range of variation, for learners, identifying the "genre signals" associated with a type of text and bringing them into alignment to produce an instance of a genre can, understandably be quite challenging. And as Diaz, et. al. (1999) have reported, as teachers, we may make things more challenging if the kinds of situations and the types of texts we solicit from students do not approach the range of variation expected in the broader discipline or community that the student writer seeks to join.

2. Looking For a Genre Signal in Citation Patterns of Academic Writers

In this article we report on a project that explores the possibilities of using computational methods to create an assistive environment for advisor-advisee mentoring in academic writing. In doing so, we hope to explore the problem space that Pare has defined and to contribute resources that may help both learners and those who are trying to help them. Our approach involves two main analytic passes. In the first pass, we use text-processing techniques that treat the texts in the corpus we are working with as strings or, more plainly, lists of words. The analytic possibilities open to us when we treat the text this way include calculating word frequencies and adjacencies. We are, in effect, looking at the words themselves as carriers of important information.

The goal of the first pass is to find genre signals - usually repetition of words or word pairings - that correspond with interesting structures such as those we might ask human raters to find in a qualitative analysis of text. We are guided in the first analytic pass by RGT which posits that generic utterances are, fundamentally, instances of repeated social action (Miller, 1994) and that, generally speaking, genre stability as indicated by regularized textual form arises from habitual responses to recurring social exigencies (Schryer, 1993).

The result of our initial analytic pass is a simple coding scheme – much less nuanced than those human raters would typically use – but that can sometimes produce results similar to human-rated texts. How does this work? Usually it is because we find a structure that is reliably present and can serve as an indicator for some broader structure or more nuanced "rhetorical move" (Swales, 1990). For instance, one of our earlier projects can be used to reliably classify discourse passages of varying lengths as being "science" or "not science" by finding instances of propositional hedging: the move to adjust claims to match the strength of available evidence that is so characteristic of scientific reasoning (Swales, 2014). The Hedge-o-Matic (Omizo & Hart-Davidson, in press) is not reading the same context cues or focusing on the same details that a human reader might to arrive at a similar conclusion. It merely looks for what we think of metaphorically as a "key protein" in the complex molecular structure of scientific discourse. When that protein is present in sufficient concentration, it makes a judgment. Hedge-o-Matic is successful because of the stability of scientific discourse as a genre – not in the formal textual sense – but in the way Bakhtin (2010) theorizes "speech genres." Without propositional hedging, one simply is not abiding the social contract scientific discourse requires.

To get to the kinds of results produced by the Hedge-o-Matic, however, another analytic step beyond the first-pass text analysis is required. In this second step, we further transform the text corpus into a graph, that is a set of nodes connected by links. Converting a text into a graph gives us a complimentary set of analytic possibilities that arise from the fact that we now have not only have words – which become nodes in the graph – but we also have edges, which represent the relationships between words. Graph analysis affords understanding not only of individual nodes and their properties, but also the holistic structure of the graph and what roles specific nodes or groups of

nodes play in it. For the Hedge-o-Matic, for instance, we can not only evaluate when a given sentence contains a hedging move, but also when it contains a hedge that is near a central claim both in terms of the graph and of meaning.

In both analytic passes, how we transform the texts for processing matters a great deal to the outcome. There is interpretive work that must go into these moves, guided not only by the general theories of genre cited above but also by more specific theories or hunches about the types of structures we are hoping to locate, isolate or amplify. The work of preparing texts for analysis, in other words, can serve as an act of theorizing rhetorical structures. Rather than writing out an exegesis, we are instead inventing analytic "recipes" to try and detect rhetorical moves of interest. For this reason, we detail our analytic method at length. We do this not only to make our methods available to others who may find them useful, but also to invite scrutiny, replication, and extension of our theoretical choices much as we would in other forms of rhetorical reasoning about texts.

3. First Analytic Pass: Searching for Citation Moves

The qualitative coding scheme described by Karatsolis in this issue used with human raters is far too complex and nuanced to use in creating a computational analysis. Right away, we needed to develop a simpler schema. Our goal was to explore the kinds of signals that might be present in the corpus that could possibly account for the human raters' interpretations of the citation types. This led, as we discuss in more detail below, to the development of an alternative qualitative coding scheme.

Like our colleagues, our analytic passes sought to focus on discursive units above the level of a single word or lexical item. We were looking for something akin to Swales' (1990) concept of "rhetorical moves" which are typically understood to be meaningful patterns of discourse within a given genre or discourse community that do important signaling work. In addition to carrying semantic meaning, rhetorical moves communicate something about a text's status as a response to a familiar kind of exigency to a particular audience, in this case readers of academic research articles. For example, one common move in scientific discourse is a claim, usually a statement of fact that is novel. Claims of fact generally signal to readers the unique contributions of the author and therefore they also invite scrutiny. Claims are typically accompanied by other moves, statements that describe evidence and as, in science writing, hedges that qualify the strength of claims based on the strength of the evidence. Our understand of rhetorical moves is shaped by Miller (1984) and others' understanding of genres as a kind social activity that, over time and as genres become more stable, is observable in the form of textual regularity.

Citation patterns are generally understood to do some of this kind of signaling work (Geisler, 1994; Prior, 2013). The key way our approach differs from that of our colleagues is that, like our previous projects using machine learning, we seek ways to

develop assistive technologies for writers and readers of text. Because of their speed and accuracy, such technologies may help more junior members of a discipline learn with and from the writing of more senior members of the same discipline.

To begin, we approached the data set with the following set of questions:

a. Can we find evidence of classifiable patterns in citation moves that contribute to genre stability; are these moves correctly and consistently identifiable with the writing cohorts from which the samples come (e.g. experienced vs. less experienced writers)?
b. Can we develop classifiable categories to produce comparisons that correspond with Karatsolis' results?
c. Can we develop analytic results that, if reliable, might be beneficial when applied to the advisor/advisee dyad for purposes such as academic mentorship?

4. Developing a Simplified Coding Scheme for Citation Analysis

Qualitative classification schemes for in-text citations have been proposed since the formation of citation analysis as a field of inquiry. These schemas, which accelerated search and retrieval, might also be treated as rhetorical because they are characterized by a motivation to persuade. Moravcsik and Murugesan (1975 p. 88) offer a four part taxonomy of in-text citations with each part divided into two valences (for an empirical application of the following categories see Cano, 1989).[1] Chubin and Moitra (1975) revise Moravcsik and Murugesan (1975) typology offering six categories of in-text citation classification in an effort to measure the impact of a science article on the development of a discipline (in this case, high energy physics). Chubin and Moitra's (1975) scheme outlines rules that govern how the claims and findings of a research paper negotiate with the wider field of scholarship, which might include the citation of related information key to the understanding of the current article and the negation of competing arguments.[2]

Cozzens (1989) offers broader axioms to understand in-text citations in scientific publications, dividing in-text citation types between communicative systems of reward and rhetoric. In the former, in-text citations are used to confer recognition on other scholars or acknowledge a type of debt paid to the intellectual property of the field (see Kaplan, 1965). In the rhetorical dimension, in-text citation types are used as a means to foster the referring paper's ratification through rhetorical devices. White and Wang (1997) offer a nine categories of classification revolving around the contribution a reference makes to the argument of the referring paper, much of which can be glossed as a mixture of in-text citation content description and/or a rhetorical move. For example, White and Wang's (1997) category of Analogy/Contrast/Comparison indicates an act of interpretation, in which the referring authors are relating their work to another. In contrast, the category of Data indicates when data from another research source is being used in the argument of the referring paper.[3] Teufel et al. (2006) approach the

classification of in-text citations as a supervised machine learning problem, that is a problem of machine learning using a set of training data. Specifically, they describe tagging in-text citations based on 12 functions, which focus on the contrastive use of references by the referring paper and the positive or neutral valence of use of the reference in the referring work (p. 104-105). They then enhanced their classification methods by also including textual features such as parts of speech and metadiscursive phrases characteristic of their research corpus (computational linguistics) in their training data.[4]

For this study, we have sorted in-text citations into 4 categories: Extractions, Groupings, and Author(s) as Actant, and Non-citations. This coding scheme is derived in large part by the lexical patterns and rhetorical uses of the in-text citations found during our initial exploration of data from the SpringerOpen Journal archive--much of which are constrained by Harvard-style citation rules and the SpringerOpen markup templates used in the presentation of research articles. In developing our coding scheme, we screen scraped 505 research articles from journals hosted by Springer OpenAccess. These journals are peer reviewed and write to the genre conventions of academic audiences, including the Introduction-Methods-Results-Discussion (IMRaD) format often used to structure scientific and social scientific journals (see Christensen and Kawakami, 2009; Hannick and Flanigan, 2013; Salager-Meyer, 1994). This screen scrape captured the meta-data of the article (author names, date of publication, institutional affiliation, and document object index), the full text of the article without images, and works cited list. Only articles types labeled as "Research Article" by the Springer OpenAccess filtering tool were used for this exploratory analysis.

The purposes behind this benchmarking task are two-fold. First, we wished to see whether or not our categories could account for every potential sentence type that may be encountered in an academic research article on a lexical level without the need for domain knowledge. The 505 research articles from Spring OpenAccess provided a corpus of academic writing that allowed us to tailor our coding scheme in ways that maximize both the syntactic and argumentative signals that each sentence in a research article could provide for computation while minimizing possible overlap or ambiguity between categories and lessening the need for domain-specific knowledge of the sampled journals. An example of a coding situation in which domain knowledge is needed involves Moravcsik and Murugesan's (1975) type 3 category, evolution of juxtaposition. A candidate citation might be obvious to a non-expert via metadiscursive clues ("Our study builds upon the work X" or "Our study differs from the work of X in the following ways"); however, where metadiscursive cues are lacking, a decision between evolutionary or juxtaposition may be more difficult. The consistent application of a type 1 perfunctory citation category used in Moravcsik and Murugesan (1975) and Cano (1989) does demand familiarity with the journal's discipline. The system of rewards described by Cozzens (1989) might also introduce conceptual ambiguity because the inclusion of a reference could both be seen rewarding a colleague and rhetorically appropriate the argument. Moreover, the notion of reward asks us to move

far beyond the lexical level and speculate at the intention of the writer, which, barring an interview, is inaccessible in this study and in most other computational approaches (similar concerns are raised in Teufel et al., 2006).

The second goal of our use of the Springer OpenAccess corpus was to also fabricate a sentence parser that would automatically cull citation and non-citational sentences and then sort these sentences into finer-grained categories based on lexical information. While we did not use this sentence parser to code the sentences from Karatsolis' data, we do look towards future applications in which such a program can undertake more sizeable text processing jobs.[5]

Our approach to a lexically-driven but still rhetorically informed coding scheme takes its rationale in part from the "shallow analyzer" method described in Marcu (1997). In his effort to parse natural language texts for their rhetorical function, Marcu (1997) compiled pre-marked cue phrases with explicit grammatical functions and rhetorical uses. In one example, the cue phrase "Although" signals both a clausal break in the text and a "concession" based on its limited uses in English. In another, the cue phrase "yet" indicates an antithesis in the argument (Marcu 1997 p. 101). In our coding scheme, we take the four different citation categories as meaningful in their lexical distinctions just as Marcu uses key words such as "although" or "yet" to delimit the possible rhetorical interpretations of a sentence. Naming or not naming an author in a sentence is taken to be meaningful within the sentence itself, even without contextual or intertextual information. Given the constraints of our coding scheme, we feel that the categorization of citations into four types can provide relevant quantitative and qualitative information about citation practices in academic writing and function as a horizon line between mentor and mentee texts.

We describe the requirements for each category and their rhetorical contributions to the analysis in turn.

4.1 Extractions

Extractions include in-text citations that present an idea from a source without involving the source in the syntactic predication of the idea. A typical example would be an idea paraphrased from source and attributed via a parenthetical reference. From Kemner et al. (2015):

> The presence of psychopathology is often explained on the basis of stress-diathesis interactions (Monroe and Simons [1991]).

In the above case, we can see a move to prioritize the information from the article by Monroe and Simons through conventional citational parentheticals as opposed to prioritizing the work of Monroe and Simons as active agents in the review of existing scholarship. Such a move could conceivably be driven by an effort to own the voice of expertise in the work or stylistic (e.g. a choice to describe conclusions as facts rather than narrating research as a process). Thus, Extraction citation would fall into the

category of "non-integral citations" discussed by Swales (1981) and Swales (2014), in which the name of a referenced author is captured in a parenthesis or as a numeric index to a works cited list.

4.2 Groupings

Groupings include in-text citation sentences that lists 3 or more sources within a parenthesis or brackets. For example from Kemmer et al. (2015):

> Interestingly, several studies have demonstrated that a history of episodes is a significant risk factor for future recurrences in mood disorders (Judd et al. [2008]; Keller et al. [1983]; Perlis et al. [2006]).

The in-text citation from Kemner, et al. (2015) above is classed as Grouping because it recounts 3 sources that have found evidence that significant life events/difficulties influence the manifestation of unipolar depression and bipolar disorders. As a rhetorical move, Grouping might resemble a gross form of "parenthetical plonking" discussed in Swales (2014) or, as Swales puts it, "nods all around to previous researchers" (p. 135). However, we treat Grouping or "plonking" maneuvers as attempts to synthesize a range of specific findings that cohere around a broader topic. We would argue that the appending of sources to the generalized idea functions as a truncated literature review, in which researchers such as Kemner et al. (2015) can demonstrate their awareness of relevant sources and establish their affiliation to a community of researchers interested in mood disorders. Another example from Kemner et al. (2015) might illustrate this more clearly:

> This has also been found in previous studies (Bender and Alloy [2011]; Hunt et al. [1992]; Kessing et al. [1998], [2004]), but it was hypothesized that this might be due to life events occurring as a consequence of the disease (Kessing et al. [2004]).

In the above case, the authors are tracing a lineage of studies related to their own project, which both corroborate and depart from their own findings.

An example from Leighton (2014) from the Journal of Evolutionary Education and Outreach employs what we are calling a Grouping in-text citation to gloss entire fields of study through parallel references:

> The maintenance of public goods has attracted researchers in biology, as well as in economics, sociology, and psychology (Hardin [1998]; Boyd et al [2003]; Bowles [2006]).

In the above case, each reference in the parenthetical list serves as representative of a wider topicality. Consequently, the Grouping categories both derives from the syntactic grouping of sources in a parenthetical list and how this list functions to situate the current research in relation to a disciplinary group. Such a move is consistent with the "create a research space" (CARs) model discussed by Swales and Najjar (1987). Swales and Najjar's (1987) 4-move model of crafting introductions for research articles also applies. In brief, Swales and Najjar argue that the genre of the research article

introduction is routinely involves (Move 1) signaling to the reader the import of the present research; (Move 2) summarizing relevant prior work in the field or on the research topic; (Move 3) indicating an insufficiency in the prior research; (Move 4) explaining how the present study will address current gaps in knowledge (1987, p 178-179). The Grouping category in our study might be considered a means to "create a research space" by presenting to the reader Swales and Najjar's Moves 2 and/or 3 within the confines of a single sentence.

Admittedly, the first principles of the Extractions and Groupings categories overlap, and both would be considered non-integral citations in Swales' (1990) formulation. A sentence which contains a summation of research and a blitz of references in a parenthetical is still, in a basic sense, evidence of an extracted idea located in the metaphorical margins of the author's argument. The rhetorical difference is one of valence dependant on a ratio of summarization. An idea distilled in a sentence from one or two sources suggests less filtering or homogenization of the source, which emphasizes what particular agents are saying in relation to the topic of present article. An idea distilled in a sentence from three or more sources would suggest a greater homogenization of source material, which emphasizes what a community of scholars is saying in relation to the topic of the present article. Another way to distinguish the moves of Extractions and Groupings is to understand them as acts of enrollment. Ultimately, every citation is a means to enroll sources into a study. For our purposes, we assess enrollment by its conspicuousness at the sentence level. A listing of three or more sources in a parenthetical draws attention to the breadth of interest in a particular topic and foregrounds the present diligence of the present authors to familiarize themselves with the voices of fellow experts. An Extractions in-text citation may do the same, but such a sentence devotes less space on the page to dramatize this type of engagement.

4.3 Author(s) as Actant(s)

Author(s) as Actant(s) refers to in-text citations that feature the authors of the cited research as clausal subjects or objects of the sentence, objects of subordination, as the originators of a direct quotation, or if the author's name is related to research or an idea through a possessive contraction. These in-text citations must contain the proper names of authors and a parenthetical date of publication reference or a bracketed numerical reference. In this way, Author(s) as Actant(s) citations can be equated to a narrow-form of the "integral citation" type identified by Swales (1990). The same syntactic pattern is also employed by Thompson and Ye (1991), who focus on "canonical" citational forms: a proper name followed by a parenthetical or bracketed date used as the subject or object in the clause (see also Charles' (2006) study of reporting verbs in citations for another cognate use of "integral citation").

The Author(s) as Actant(s) category ignores pronoun attributions and/or those attributions that lack a date of publication. For example, a sentence pattern that elaborates on a previous in-text citation such as "They further argue that . . ." would not

be considered an Author(s) as Actant(s) citation in this coding scheme; it would be categorized as a non-citation. Self citations that use the pronoun "I" or "we" that have an appearance of an integral citation are classed as an Extraction. References in which a source is introduced through advisement such as "(see Smith, 2007) would not be considered an Author(s) as Actant(s) citation type because it is subordinated within the sentence through a parenthesis. However, if the recommendation to "see" an author is integrated into the main argument of the text, that sentence is classed as an Author(s) as Actant(s).

An example from Keown-Stoneman et al. (2015) illustrates a case in which the author of a work being cited acts as the subject of the sentence:

> Duffy et al. ([2010]) suggested that specific types of psychopathological manifestations are precursors for bipolar disorder in this high-risk population.

In the above example, "Duffy et al." is the subject of the sentence in which they "suggested" an idea about possible precursors to bipolar disorders. In such a case, we consider the research space being created by Keown-Stoneman et al. (2015) as one that includes a diegetic dimension in which cited authors act or are acted upon at the sentence level.

An example of an Author as Actants in-text citation in which the cited author is acted upon as the object of the predicate can also be found in Keown-Stoneman et al. (2015):

> A more detailed description of the collection methods can be found in Duffy et al. ([2007]).

In the above example, the Keown-Stoneman et al. (2015) denote the existence of additional research that has been written about by Duffy et al., which implies that work has been done by Duffy et al. In one sense, the above in-text citation is the passive voice inverse of "Duffy et al. ([2007]) offer a more detailed description of these methods."

An example of an Author(s) as Actant(s) in-text citation in which the named authors appear as objects of actions in a subordinate clause is illustrated by the following sentence from Correa Bahnsen, et al. (2015):

> Subsequently, we evaluate the cost-sensitive logistic regression (CSLR), estimated using the default parameters as suggested in (Correa Bahnsen et al [2014a]).

In the above case, the Correa Bahnsen, et al. (2015) are gesturing back to a previous publication, in which Correa Bahnsen, et al. have performed the action of suggesting methods for a cost-sensitive logistic regression. In the above case, the parentheses are ignored as an artifact of the journal markup template, and the named authors are treated as the object of the preposition "in."

For Thompson and Ye (1991) the naming of an author in conjunction with a reporting verb such as "show" or "confirm" or "provide" represent acts of interpretation, which

signal the writer's orientation to the referenced material. This orientation could involve agreement or disagreement between positions or facts or function as a mechanism for reward, validating the achievements of the cited author (Charles, 2006 p. 322). Paul (2000) reasons that the decision to name the author(s) of a reference is reflective of that source's centrality in the field; otherwise, the reference could be glossed numerically or within a parenthetical (p. 199). Identifying an author by name is an act of spotlighting that reference. Thus, for Thompson and Ye (1991), Paul (2000), and Charles (2006), instances of nomination are built around a kernel of conspicuous evaluation. The form that these evaluations take are dependent upon more textual factors then our coding scheme allows. For example, a negative, positive, or neutral evaluation would need to account for the type of reporting verb used (Thompson and Ye, 1991 p. 372). The citations "We oppose Smith (2000)" and "We agree with Smith's (2001) findings" would turn the polarized valences of "oppose" and "agree." However, our coding scheme only accounts for the operation of a verb in conjunction with a named author, not the type of verb. What we are interested in is how the act of naming an author as the subject or object of an action enlarges the context of the reference to include specific practitioners in the field. This rhetorical move then makes it possible for writers to affirm, extend, complicate, or challenge related work, all of which makes possible a qualitatively different means to engage with sourced material, and, by extension, a fashion a different type of rhetoric than that offered by an Extraction or Grouping citation type.

4.4 Non-citations

All sentences that do not fit with the Extraction, Grouping, and Author(s) as Actant(s) citations would be placed in the category of Non-citations. This includes sentences that might be considered references in other taxonomies, such as attributions that replace the named agent with a pronoun and lack a parenthetical date or a parenthetical with a name and date or a bracketed numeral reference.

Given the above description of our coding scheme, it would be fair to say that our categories do elide traditional attribution activity as non-citations, especially if the attribution activity involves an elaboration of a previous citation. Part of this elision is a concession to the limitations of the shallow parsing program we used and an effort to minimize subjective coding decisions. Thus, the type of citation activity we are attempting to capture through shallow parsing and the formulation of network graphs might be more accurately described as a measure of citational intrusion whereupon authors are making manifest their adherence to research conventions and signalling adjuncts to their arguments. If we consider our citational categories as rhetorical gestures in the vein of Gilbert Austin or Francoise DelSarte then an Extraction might be equated to a nod; Grouping, a rounded sweep of the hand; and Author(s) as Actant(s) as pointing.

5. Analytic Pass 2: Seeing Citations as Important Structures in Graphs

Our analysis is based on a network graph structure, which draws upon the citation coding scheme decisions for its configuration using the numerical codes of 0-no citation, 1-extraction, 2-grouping, and 3-author as actant for a node list of [0, 1, 2, 3]. Thus, the network graph is comprised of only these four nodes. Edges are drawn between one node and its next immediate neighbor. The sequence of nodes is determined by the original arrangement in the source text. For example, the follow passage from Lemieux (2015) has been tagged in our SpringerOpen excavation in the following manner:

> (Digital photographs with spatial information are commonly referred to as geotagged photos, 0)
>
> (Geotagged photos are created in a variety of ways categorized as manual or automatic (Welsh et al. [2012]), 1).
>
> (Automatic geotagging is possible using digital cameras with a built-in or connected GPS, 0)
>
> (Smartphones are an emerging system with a built-in GPS receiver (Valli and Hannay [2010]) however many camera companies (i.e. Casio, Nikon, Panasonic, Olympus) also sell devices with this feature., 1)

The citation tags would results in the following edge relationships between the sentences: [(0-1), (1-0), (0-1)]. Because the tags are proceeding in one direction, we could also represent the preceding list of edges as a path that connects 0-1-0-1. Because the citation coding scheme is collapsing distinctions among unique sentences and rendering each sentence in terms of a class, it is more than likely that each graph will feature nodes that link to themselves (e.g. 0-0). The graph representation suggests an identity among nodes with the same class assignment, even though none is meant to exist. To compensate for this elision, multiple edges are drawn between nodes, duplicating the arrangement of the annotated source text. Consequently, this graph structure contains both direction and multiple edges between nodes with self-loops. In conceptual terms, the graph structure is an Eulerian path--a graph in which all vertices have an equal amount of incoming and outgoing links save for the first and final vertices despite the ostensible presence of repeated edges and self-loops. This graph structure is equivalent to a network graph in which each node contains a class value and a unique identifier. In such a case, each node and edge would be unique. The reason we have adopted the network graph structure of an Eulerian path is because we wished to preserve the sequencing of the natural language text which involves (in English) the linear progression of sentences. By representing the text as an Eulerian path, we maintain the sequence of the citational and non-citational moves. As discussed below, creating an Eulerian path will allows us to extract features that writers

make use of by default, including organizing textual content through a beginning, middle, and end.

The resulting text-to-graph structure allows us to perform computational operations on an array of coding decisions. These operations generate a series of features from each graph structure that then allows us to construct a model for comparison between advisor and advisee texts. In the following section, we describe the features that we extract from the text-to-graph structure and the rationale for using such feature for a rhetorical analysis of in-text citation practices.

6. Graph Features As Indicators

6.1 Eigenvalues of the citation nodes of the graph adjacency matrix

This feature comprises of the eigenvalues of the three citational node types in the graph (i.e., 1, 2, 3). These values are computed by taking the highest value of the eigenvector of the adjacency matrix of the graph. This highest value is the eigenvalue of the eigenvector, and is considered the characteristic or basic value of the matrix. For our analysis of advisor and advisee texts, we use these eigenvalues as a baseline for comparison because they refer to essential aspects of the graph's structure. Because the graph structures we have created contain self-loops, the adjacency matrices of the graphs will often be asymmetrical with values disproportionately skewed toward edges between non-citational nodes (0-0 or sentences without in-text citations that follow other sentences without in-text citations).

Given that our main interest in graphing advisor and advisee texts is in citational practices, we only retain the eigenvalues in the adjacency matrix that govern citational nodes (1, 2, 3). In a sense, we already know that the non-citational sentences constitute the majority of the sentence count in advisor and advisee texts based on the generic nature of academic writing. Consequently, it is not revelatory to see that the 0-node vector possess high eigenvalues or eigenvalues that largely condition the nature of the graph. As such, we are more interested in the limited but strategic use of citations within the larger network of non-citations.

6.2 Subgraph size range

This feature consists in removing edges from the graph that contain a 1, 2, or 3 node as its inbound or outbound link. The result is an Eulerian trail that is segmented at the points of deletion. For example, with the deletion of all 1-citational edges, the Eulerian trail with the following path of 0-0-0-0-0-1-0-0-0-0-1-0-0-2, would be reconfigured as three separate sequences or subgraphs: 0-0-0-0-0; 0-0-0-0; 0-0-2. We view this feature as similar to the notion of network robustness (Albert, et al., 2000), which is a measure of how well a network can withstand node deletion while retain its connectivity. The findings of Albert, et al's (2000) study of network robustness is not directly transferable to the types of graph structures created in this study. First, Albert, et al. (2000) are

concerned with comparing the properties of exponential and scale-free networks when attacked or confronted with node malfunctioning. These scale-free networks are organized by preferential attachment, and contain a network structure characterized by a few key nodes with many edges and a prevalence of nodes with significantly less edges. If we take the nodes of our graph structure as representative of the classes 0, 1, 2, 3 and chart the counts of edges, we do see modest support for a powerlaw distribution within each article of the Springer OpenAccess corpus and in the advisor and advisee texts when compared with an exponential fit. However, a similar case for support could be made for a lognormal distribution over a powerlaw distribution. Consequently, we cannot attribute the same qualities of robustness described in Albert, et al. (2000) to the Eulerian paths employed in our analysis.

Second, the objects of study in Albert, et al (2000) consist of websites on the World Wide Web, conceived of as nodes in an enormous directed graph. Websites are not citational and non-citational sentences within a single article, but the operant difference is sequentiality. While a cluster of local sports websites might link to espn.com, they are free to make additional connections to other websites and be the recipients of other incoming links. Because the nodes in our graph structures represent unique sentences in a text, they are fixed in a linear sequence. This challenges the application of Albert, et al's (2000) notion of network robustness in the following crucial ways: where the random deletion of a node in a scale-free or exponential network may affect network connectivity, it is still possible that information from one node could reach another via alternative paths. In the Eulerian paths that we are using as models of for citational practices, the deletion of any one node or edge will lead to short in the network. Moreover, in an Eulerian path, each node contributes the same amount of connectivity (save the first and last node in the path) because each nodes has one incoming edge and one outgoing edge. This lack of redundancy and alternate paths means that deleting any node or edge will inhibit information from passing through the network in roughly the same way but at different locations. Thus, a notion of robustness or disruption for scale-free or lognormally distributed networks does not fit our graphical models.

We base our measurement of robustness on the relative size range of the subgraphs created by the deletion of edges containing a citational node. This involves dividing the sizes of each subgraphs into quartiles and taking the difference of the first and third quartiles. In doing so, we are effectively trying to extrapolate the typical size of a subgraph fragment as a means to either infer the amount of disruption caused by citational edge deletion or the relative size of the components citational edges serve to connect in the Eulerian network. We use interquartile range as opposed to a mean of sizes to account for general skewness of subgraph sizes. The larger the subgraph interquartile size range, the larger the subgraph fragments tend to be after edge deletion, suggesting that the presence of a citational type is less frequency and widely spread throughout the text. The smaller the subgraph interquartile size range, the smaller the subgraph fragments tend to be after edge deletion, suggesting that the

presence of the citational type is more frequent and shares more proximal relationships with like citation types throughout the text.

6.3 Location and distribution of citational edges

This feature attempts to account for the position of citational edges within the network graph. The process involves dividing the network graph sequence into percentile ranges: 0-30 indicates the beginning of a network sequence or research article; 30-75 indicates the middle of the network sequence or research article; 75-100 indicates the ending of the network sequence or research article. Each citation type is counted within each percentile range and then divided by the total count of that citation type, resulting in the relative frequency of the citation type per section. This method of feature extraction replicates a study of citation use undertaken by Cano (1989). While authors of research articles are free to insert citations at any point in the text (and do so), work by Cano (1989), Voos and Dagaev (1976), and Ding, et al. (2013), Hu, Chen, and Liu (2013) (see also Tang and Safer, 2008) have found that citations in citations in the scientific research article is primarily located early in the document--within the Introduction, Literature Review, and Methods sections. In their corpus survey of the Journal of Infometrics, Hu, Chen, and Liu (2013) notes that more than half of the in-text citations in each article occurred in the first 30% of the document (p. 891). Paul's (2000) study of scientific articles in the field of chaos theory also finds that 50% of citation appear in the Introduction section of the article.[6] In our benchmark testing of the 505 research articles from the SpringerOpen database, we found a similar concentration of the citations. Discounting citational type, 46.8% of citational edges appear in the first 30% of our network sequence.[7] Consequently, we view this feature as a means to index the genre fidelity of a writer's citational practices. In general, we would expect to see a higher density of citations in the beginning of the network sequence. The lack of such a density could then indicate a departure from genre conventions. Moreover, tracking where an Extraction, Grouping, or Author(s) as Actant(s) citations congregates within a text can provide additional information about the specific generic constraints placed on each type.

6.4 Edge reciprocity

Reciprocity in directed graphs refers to a condition in which two vertices point to each other in a loop of edges. For example, node 0 points to 1 and 1 points to 0 (Newman, 2010 p. 204). In all practicality, each node in our Eulerian path is unique because each nodes represents a unique sentence in the source text. For this reason, when we speak of reciprocity, we are referring to node classes that point to each other in the linear arrangement of the graph. Thus, the nodes of classes 0 and 1 would be considered reciprocal if they follow the sequence of 0-1-0. We express edge reciprocity as the fraction of reciprocal edges to the total amount of edges. As in the case for the eigenvalues of the adjacency matrix of the graph, we are only interested in those edges formed when one citational node class links with another, understanding that there will

be a high rate of non-citational nodes linking other non-citational nodes in the network. Thus, we screen out the edge type 0-0. Each of the remaining permutations of edge reciprocity is captured as a distinct value (see Figure 1).

Edge Type
(1-1)
(1,2) and (2,1)
(1,3) and (3,1)
(2,2)
(2,3) and (3,2)
(3,3)

Figure 1: Edge Reciprocity Types.

We see this feature as a means to see how the citational types of Extraction, Grouping, and Author(s) as Actant(s) tend to mutually inform each other due to their function in the text. From a rhetorical standpoint, we might expect that there would be higher reciprocity between Extraction and other Extraction types and Extraction and Grouping types. The premise would be that both Extraction and Grouping citation types function to deracinate ideas from the referred to source. To extend Voloshinov, Extraction and Grouping citations are more assimilated stylistically and compositionally to the argument and presented as pure information. In contrast, Author(s) as Actant(s) citations would fall into the category of reported or quasi-direct speech, in which information or actions are more explicitly associated with exogenous circumstances (Voloshinov 1973, p 117), and bear lexical markings that would be differentiated from sentences that focus on the "what" rather than the the "what" and the "how (Voloshinov 1973, p. 119). In the context of Latour and Woolgar (1976), the distinction between Extraction, Grouping, and Author(s) as Actant(s) citations might be aligned along a spectrum of "facticity" in which a parenthetical references calls less attention than the clausal incorporation of an author(s) name to the interventions of other research efforts, thus, rendering the sentence more fact-like in its surface appearance (p. 80-81). In another sense, the flattening out of the reference into an Extraction or Grouping move may be more rhetorically convenient (see Cozzens 1989 p. 443) because the priority of the reference is to present uncontroversial information or acknowledge the existence of related work, thereby rendering the need to present context or qualification unnecessary both conceptually and stylistically. In our benchmarking efforts, we find this to be the case within the 500 research articles in the SpringerOpen database. The highest mean edge reciprocity score is between Extraction to Extraction nodes (1-1). The second highest mean edge reciprocity score is between Extraction to Grouping and Grouping to Extraction (1-2, 2-1). The third highest mean

edge reciprocity score is between Extraction and Author(s) as Actant(s) and Author as Actant(s) to Extraction (1-3, 3-1).

The above features are aggregated into an array in the order shown in Figure 2:

Feature
Adjacency Matrix Eigenvalue Extraction
Adjacency Matrix Eigenvalue Grouping
Adjacency Matrix Eigenvalue Author(s) as Actant(s)
Subgraph size (per quartile range) Extraction
Subgraph size (per quartile range) Grouping
Subgraph size (per quartile range) Author(s) as Actant(s)
Extraction Edge location (0-30 percentile)
Extraction Edge location (30-75 percentile)
Extraction Edge location (75-100 percentile)
Grouping Edge location (0-30 percentile)
Grouping Edge location (30-75 percentile)
Grouping Edge location (75-100 percentile)
Author(s) as Actant(s) Edge location (0-30 percentile)
Author(s) as Actant(s) Edge location (30-75 percentile)
Author(s) as Actant(s) Edge location (75-100 percentile)
(1-1) Edge reciprocity
(1,2) and (2,1) Edge reciprocity
(1,3) and (3,1) edge reciprocity
(2,2) edge reciprocity
(2,3) and (3,2) edge reciprocity
(3,3) edge reciprocity

Figure 2: Graph Feature Array.

The graph feature values tabulated in the array are then used to calculate the Euclidean distance between network graphs.[8]

7. Discussion: Citation Moves Appear Stable Enough to Reliably Locate & Classify

We return now to our three framing questions to discuss the results of our two analytic passes. We want to reiterate that all of these results are preliminary and much work remains to further validate and test for reliability the work we have done. The first qualification we should make relates to the size of the corpus reviewed; our corpus was small by big data or text analysis standards. Moreover, the texts are sparsely populated with the phenomena examined. For both reasons, we insist on calling this work exploratory versus confirmatory or even descriptive.

With that hedge in place, we are encouraged by what we have seen and believe that further research in this area is warranted. We find evidence of the genre signals that correspond with the citational patterns Karatsolis (this issue) found. Confirmation studies with larger sets of data will be needed to producing more reliable results in the future. But for our exploratory research questions, we found useful results.

1) Can we find evidence of classifiable patterns in citation moves that contribute to genre stability and may be the basis for similarities or differences in the writing cohorts from which the samples come (e.g. experienced vs. less experienced writers)?
The answer to this first questions appears to be yes. For this study, we compared articles from two cohorts of advisor and advisee texts from the Karatsolis dataset. The cohorts represent three different disciplines: chemistry advisor and advisee text and materials science advisor and advisee texts. Because the dataset was small, we hand coded citation patterns based on the coding scheme we developed from the first analytic pass, which was aided by the original annotations of Karatsolis. In this section we represent the distances between the network graphs in the distance matrices shown in Figures 3 and 4.

	CA_1	CA_2	CA_3	CAE_1	CAE_2	CAE_3
CA_1	0	5.990	3.147	2.808	7.663	4.082
CA_2	5.990	0	3.352	4.046	9.107	3.796
CA_3	3.147	3.352	0	1.050	8.288	1.761
CAE_1	2.808	4.046	1.050	0	8.005	2.439
CAE_2	7.663	9.107	8.288	8.005	0	9.738
CAE_3	4.082	3.796	1.761	2.439	9.738	0

Figure 3: Pairwise similarity scores for chemistry cohort.

The chemistry cohort contains 1 advisor, who is the author of 3 texts: CA_1 and CA_2 and CA_3. The chemistry cohort also contains 1 advisee, who is the author of 3 texts: CAE_1, CAE_2, and CAE_3. The distances separating the samples are depicted in Figure 3.

The material science cohort contains 1 advisor, who is the author of the following texts: MA_1, MA_2, MA_3, MA_4. The material science cohort contains 1 advisee, who is the author of the following texts: MAE_1, MAE_2, MAE_3, MAE_4. The distance separating all samples of the material science cohort is depicted in Figure 4.

We do see this technique as providing a clear means to compare across and within cohort texts as we hoped. And so we can now turn to the next question - do these apparent differences tell us anything interesting?

	MA_1	MA_2	MA_3	MA_4	MAE_1	MAE_2	MAE_3	MAE_4
MA_1	0	48.569	7.262	6.497	7.251	4.982	7.079	5.346
MA_2	48.568	0	46.261	46.793	47.671	48.389	46.77	47.948
MA_3	7.262	46.261	0	2.274	3.114	3.493	2.117	2.999
MA_4	6.497	46.793	2.274	0	1.7789	2.303	1.360	1.718
MAE_1	7.251	47.671	3.114	1.779	0	2.764	1.584	2.307
MAE_2	4.982	48.389	3.493	2.303	2.764	0	2.851	1.033
MAE_3	7.079	46.767	2.117	1.360	1.584	2.851	0	2.173
MAE_4	5.346	47.948	2.999	1.718	2.307	1.033	2.173	0

Figure 4: Pairwise similarity scores for material science cohort.

2) Can we develop classifiable categories to produce comparisons that correspond with Karatsolis' results?

The answer to this question appears to be a somewhat more qualified yes. We can begin by noting the average similarity scores across individual writers' texts. The average pairwise distance between the chemistry advisor's texts (CA_1, CA_2, CA_3) is 4.163. The average pairwise distance between the chemistry advisee's texts (CAE_1, CAE_2, CAE_3) is 6.72700061. For the material science pair, the average pairwise distances between the material science advisor's texts (MA_1, MA_2, MA_3, MA_4) is 26.276. The average pairwise distance between the material science advisee's texts (MAE_1, MAE_2, MAE_3, MAE_4) is 2.119. These average pairwise distances across individual writer's texts seems to follow some expected patterns. For example, we would expect that the more novice advisee texts would show less range in citation patterns than advisors texts due to the fact that (1) advisees are less experienced writers in comparison to advisors, (2) less experienced writers often have less opportunities to perform in various genres[9], (3) less experienced writers are afforded less flexibility in

their writing styles (see Berkenkotter and Huckin, 1995, p. 117-144). Consequently, we might expect that advisees hew closer to a more regular citation pattern. We see this in the chemistry cohort advisee. Among the sampled texts, the chemistry advisee primarily uses Extraction citations. Moreover, the chemistry advisee texts focus the use of Extraction citations in the early third of the papers. Conversely, the chemistry advisee primarily uses Author(s) as Actant(s) citation towards the final third of the paper where the writer discusses the results of a study or experiment. These tendencies are consistent among the chemistry advisee's writings, and are also consistent with the generic demands of a scientific research paper based on an IMaRD structure. In such a paper, the methods section, which falls early in the paper, is conceived as more denotative. Thus, citations are more likely to be Extractions meant to quickly establish the precedent of a procedure. The closing discussion section generally requires a more conspicuous engagement with previous research as a means to interpret results and mark the significance of the results to the field.

The material science cohort advisee demonstrates an even greater homogeneity, typified by the high incidence of non-citation moves and a prioritizing of citational moves in the first third of the writings. The material science advisee evidences a tendency to use citational moves as a means to create a research space (in the Swales sense) and then append new work to this rhetorical situation.

We might also expect that the distances between advisor texts to be more variable due to the fact that advisors have (1) greater disciplinary writing experience, (2) access to more genres, (3) and more flexibility to deviate from convention due to their enhanced status. We see this in a dramatic fashion with the material science advisor texts. MA_2 text is far different from any other text among the material science advisor and advisee texts. This can be attributed to the fact that MA_2's text only signals 4 citations total moves in our coding scheme. The text of MAE_2, which is of similar length in terms of sentence count (MAE_4 = 32 vs MA_2 = 54) has 9 citational moves. Moreover, the citational moves of MA_2 only occur in the first quartile of the sequence. In contrast, the citational moves found in MA_1, MA_2, and MA_4 occur with a wider distribution across each respective sequence, suggesting texts that are more dependent upon citational moves to construct an argument and illustrate results. In support of this macro-assessment, a closer look at the text of the MA_2 reveals an emphasis on reporting the experimental results of an experiment on the nanoparticle dynamics. This is indicated by the first sentence of the text, which contains a Grouping citation:

> It is well known that nanoparticles of inorganic mate-rials exhibit many properties that are unexpected from the conventional point of view (1-3)

Rhetorically-speaking, MA_2's work is meant to reaffirm convention thinking on nanoparticles. Thus, it is plausible to assume that less citational work is needed. In contrast, MAE_4 attempts to challenge existing understanding, as demonstrated by the following goal statement:

> We present here α modified CVD method derived from our carbon nanotube growth procedure-capable to grow parallel carbon fibers and extended, ordered networks of multiwalled nanotubes forming layered multiple junctions

That said, and discounting the extreme pairwise distance figure of the MA_2 text, we can discern close proximity between the material science advisor and advisee text samples. As can be seen in Figure 4, the MAE_1 text is close in Euclidean distance to the MA_4 text.

Furthermore, when we look at the texts themselves, the differences that Karatsolis notes as significant between advisors and advisees do offer some possible targets for distinguishing between the two. One key difference is the evaluation of sources. Here are two example citations that contain references. The first is from the chemistry advisee (CAE_2), the second from an advisor (CA_2). Note the evaluative language in the advisee's sentence:

> Advisee:
>
> Cox and Pilcher list an alternate measurement obtained by Long and Norrish31 (-136 ± 25 kcal mol-1) that is actually in better agreement with the calculation.
>
> Advisor
>
> Meline et al. (3) used proportional-derivative and minimum variance adaptive control to overcome the learning periods associated with adaptive controllers.

Both of these citations use the Actor-as-Actant citation pattern, but the advisor's is more descriptive, casting the work of the named authors as something akin to problem solving rather than being better or more accurate. Words like "actually" and "better" stand out in the advisee's text as possible markers that are generally absent in the advisor's text, despite using the same elaboration pattern.

This example is one of very few contrasts that stood out as noticeable at the sentence level. Most of the differences are more subtle (as the distances above indicate) and when reading the texts, they are noticeable only after coding and doing comparisons across hundreds of sentences. We would suggest where texts approach one another in similarity, it is not simply because they contain sentences that are constructed the same way. Rather, the full texts are constructed in similar ways at the macro level.

At this level, an analysis of the macro structure can be useful for showing where a writer is constructing a text that is significantly different from other writers. As seen in the similarity scores table, above, the chemistry advisee's text is quite different than that of all of the other writers, advisors and advisees alike. We can therefore suspect that there is something this writer does not know about the ways others in the disciplinary area use citations.

As Berkenkotter and Huckin (1995) have shown, proposing novel methods or results in scientific writing often goes hand in hand with citational moves because

writers, in order to establish the newness of their findings, must position their work in contradistinction to existing scholarship via citations. While the discovery of novelty is not yet the domain of the graph methods presented in this paper, the macro-perspective that the graph analytic does afford may focus research questions to a more delimited range. As in the case of MA_2, research could identify the outlier text, examine the key features of the text, and gain a general insight of what deviations to look for before approaching the remaining texts in the cohort.

Perhaps the greatest value of our approach lies not in precisely replicating each human-coded judgement using a scheme like Karatsolis'. Rather, this method may help advisors and advisees notice patterns at the macroscale, but focused nonetheless on a particular set of rhetorical moves (e.g. citations and their use) in order to better raise awareness of the need to do more targeted reading, commenting, and revision. In this way, automated analysis can become a tool to help those guiding learners to understand the range of rhetorical practices that novice writers command and to evaluate whether these are inclusive of a repertoire of moves appropriate for a given disciplinary community.

3) Can we develop analytic results that, if reliable, might be beneficial when applied to the advisor/advisee dyad for purposes such as mentoring academic writing?

This remains to be seen, but we have some reason to be optimistic based on the results we see here. If we begin with Karatsolis' finding that advisees do much more elaboration than advisors, we see an opportunities to detect these and - perhaps - flag them for consideration during the writing and revision stage.

One of the most promising opportunities we see is based on possible correlations in citation types and Karatsolis' findings. For instance, we noticed that the author as actant category of citation may be a more efficient way to find the pattern of more/less elaboration around citations (i.e. author as actant moves mean there are more likely to be elaboration). Here is an example taken from one of the chemical engineering advisee's texts:

> Felinger and Guiochon [20] employed a modified simplex algorithm to carry out the optimization of experimental conditions in displacement systems. However, they employed the equilibrium-dispersive model that restricts their results to stationary phase materials with relatively small particle sizes (5–20m).

There are two good reasons to expect more elaboration from an author as actant move like this. First, as Paul (2000) argues, the choice of an author as actant citation type is a more demanding option (when compared to an Extraction or non-integral citation type), and signals a move to engage with the specific works of other researchers. In making a conspicuous point of this engagement, it is reasonable to also expect that whatever points being offered will often exceed the bounds of the author as actant sentence as it does in the example above. Thus, it is possible that we can view author as actant

citational types as a indicators for elaboration, although more extensive text mining work would need to be conducted on a larger corpus to verify such expectations.

We also believe that the comparative figures yielded by the measurement of pairwise distances can help establish a baseline of fidelity between advisor and advisee writings that may shed light on how advisees are acquiring citation practices and how advisors are modeling citation practices. Mastery of genre conventions is and will always be a moving target. While the lack of expected citational moves may inhibit publication, a strict adherence to the quantity and distribution of ratified citational moves may not lead to persuasive findings or it may stifle innovation. What our network graph metrics offer is a means to automate the discovery of a generic baseline for citational moves among academic mentoring relationships. For example, once a qualitative judgment about the acquisition of disciplinary writing codes has been established by an advisor, a pairwise metric calculated in the future can serve as a global indicator that an advisee is maintaining the proper citational patterns that allows for field recognition and the preservation of that advisee's scholarly voice. Beyond the mentoring relationship of advisor and advisee, we also foresee an ability to compare citational moves across other disciplinary aggregates such as journals, dissertations, and scholarly monographs in order to examine how a field is being constituted via citations.

Notes

1. Moravcsik and Murugesan's (1975) 4 part citation coding scheme includes the following:
    a. conceptual (citation references a theory or idea) or operational (citation references a method)
    b. organic (citation necessary to the understanding of the content of the current paper or referenced paper) or perfunctory (acknowledgement of previous work)
    c. evolutionary (current paper building on work done in reference) or juxtapositional (current works provides an alternative to referenced work)
    d. confirmative (current paper affirms work of referenced work) or negational (current paper challenges or critiques referenced work)
2. Chubin and Moitra (1975, p. 426-427) in full include:
    a. essential basic (citation is integral to the content of the referenced article)
    b. essential subsidiary (referenced work or findings are integral to understanding the referenced work but not related to the content of the referring paper)
    c. supplementary (references provide additional, independent information)
    d. perfunctory (references included with interpretation)
    e. negational partial (references that the referring article disagrees with in part)
    f. negational total (references that the referring article rejects outright).
3. For a review of citation classification taxonomies, see Cronin 1984, p. 35-49
4. The work presented in Teufel et al. (2006) builds on earlier classification efforts that seeks to automatically diagnose the content of research articles by extracting information about the rhetorical status of sentences. The assignment of rhetorical status on a sentence per sentence

basis arguably allows for a better sense of context, leading to the possibility of improved automatic text summarization (Teufel and Moens, 2002) and more informative citation indexing (Teufel 2006).

5. The text processing routine developed for categorizing citational and non-citational sentences in the Springer OpenAccess research database relies on the application of predetermined regular expression rules. The complete process is described here: http://ryan-omizo.com/2016/01/11/finding-genre-signals-in-academic-writing-benchmarking-method/
6. Paul's (2000) study also incorporates location as an operant citational feature.
7. 30.2% of citations appear in the next 30-75 percentile; 22.6% of citation appear in the final 75-100 percentile.
8. For this study, we employ sci-kit learn's pairwise distance metrics algorithm (see Pedregosa, et al. (2011) and scikit learn's module documentation at http://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.pairwise_distances.html)
9. Swales (1990, p. 213) suggests that in order to acclimate students to the rhetorical demands of academic writing, that models be presented in "caricature" format, which "simplifies and distorts" genre features. We see this process of attenuation as a means to constrict the variability of student writing.

References

Bakhtin, M. M. (2010). *Speech genres and other late essays*. University of Texas Press.

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science, 40*(4), 284-290. doi: 10.1002/(SICI)1097-4571(198907)40:4<284::AID-ASI10>3.0.CO;2-Z

Charles, M. (2006). Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines. *English for Specific Purposes, 25*(3), 310-331. doi:10.1016/j.esp.2005.05.003

Christensen, N. B., & Kawakami, S. (2009). How to structure research papers. *International journal of Urology, 16*(4), 354-355. doi: 10.1111/j.1442-2042.2009.02278.x

Chubin, D. E., & Moitra, S. D. (1975). Content analysis of references: adjunct or alternative to citation counting? *Social studies of science, 5(*4), 423-441. doi: 10.1177/030631277 500500403

Correa Bahnsen, A., Aouada, D., & Ottersten, B. *A novel cost-sensitive framework for customer churn predictive modeling*. Decision Analytics.

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

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics, 7*(3), 583-592. doi: 10.1016/ j.joi.2013.03.003

Dias, P., Freedman, A., Medway, P., & Par, A. (2013). *Worlds apart: Acting and writing in academic and workplace contexts*. Routledge.

Geisler, C. (1994). *Academic literacy and the nature of expertise: Reading, writing, and knowing in academic philosophy*. Routledge.

Hannick, J. H., & Flanigan, R. C. (2013). How to prepare and present scientific manuscripts in English. *International Journal of Urology, 20*(2), 136-139. doi: 10.1111/iju.12041

Hu, Z., Chen, C., & Liu, Z. (2013). Where are citations located in the body of scientific articles? A study of the distributions of citation locations. *Journal of Informetrics, 7*(4), 887-896. doi: 10.1016/j.joi.2013.08.005

Kemner, S. M., van Haren, N. E., Bootsman, F., Eijkemans, M. J., Vonk, R., van der Schot, A. C., ... & Hillegers, M. H. (2015). The influence of life events on first and recurrent admissions in

bipolar disorder. *International journal of bipolar disorders, 3*(1), 6. doi: 10.1186/s40345-015-0022-4

Keown-Stoneman, C. D., Horrocks, J., Darlington, G. A., Goodday, S., Grof, P., & Duffy, A. (2015). Multi-state models for investigating possible stages leading to bipolar disorder. *International Journal of Bipolar Disorders, 3*(1), 5. doi: 10.1186/s40345-014-0019-4

Lemieux, A. M. (2015). Geotagged photos: a useful tool for criminological research? *Crime Science, 4*(1), 1-11. doi:10.1186/s40163-015-0017-6

Marcu, D. (1997, July). The rhetorical parsing of natural language texts. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics* (pp. 96-103). Association for Computational Linguistics. doi: 10.3115/979617.979630

Miller, C. R. (1984). Genre as social action. *Quarterly journal of speech, 70*(2), 151-167.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social studies of science, 5*(1), 86-92. Retrieved from http://www.jstor.org/stable/284636

Ogada, M. J., Mwabu, G., & Muchai, D. (2014). Farm technology adoption in Kenya: a simultaneous estimation of inorganic fertilizer and improved maize variety adoption decisions. *Agricultural and Food Economics, 2*(1), 1-18. doi: 10.1186/s40100-014-0012-3

Otten, S., Spruit, M., & Helms, R. (2015). Towards decision analytics in product portfolio management. *Decision Analytics, 2*(1), 1-25.

Paré, Anthony. (2014). Rhetorical genre theory and academic literacy." *Journal of Academic Language and Learning 8(*1), A83-A94.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825-2830.

Prior, P. (2013). *Writing/disciplinarity: A sociohistoric account of literate activity in the academy*. Routledge.

Rissler, L. J., Duncan, S. I., & Caruso, N. M. (2014). The relative importance of religion and education on university students' views of evolution in the Deep South and state science standards across the United States. *Evolution: Education and Outreach, 7*(1), 1-17. doi: 10.1186/s12052-014-0024-1

Salager-Meyer, F. (1994). Hedges and textual communicative function in medical English written discourse. *English for specific purposes, 13*(2), 149-170. doi: 10.1016/0889-4906(94)90013-2

Schryer, C. F. (1993). Records as genre. *Written Communication, 10*(2), 200-234. doi: 10.1177/0741088393010002003

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Swales, J. (2014). Variation in citational practice in a corpus of student biology papers from parenthetical plonking to intertextual storytelling. *Written Communication, 31*(1), 118-141. DOI: 10.1177/0741088313515166.

Teufel, S. (2006). Argumentative zoning for improved citation indexing. In *Computing Attitude and Affect in Text: Theory and Applications* (pp. 159-169). Springer Netherlands. doi: 10.1007/1-4020-4102-0_13

Teufel, S., & Moens, M. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics, 28*(4), 409-445. doi: 10.1162/089120102762671936

Teufel, S., Siddharthan, A., & Tidhar, D. (2006, July). Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 103-110). Association for Computational Linguistics.doi: 10.3115/1610075.1610091

Thompson, G., & Yiyun, Y. (1991). Evaluation in the reporting verbs used in academic papers. *Applied linguistics, 12*(4), 365-382. doi: 10.1093/applin/12.4.365

Voos, H., & Dagaev, K. S. (1976). Are All Citations Equal? Or, Did We Op. Cit. Your Idem?. *Journal of Academic Librarianship, 1*(6), 19-21.