7

Perceiving Genre with Special Reference to the Academic Writing

Dr. Karansinh Rathod

Associate Professor, Dept. of English, DKV Arts College, Jamnagar

Abstract:

This study uses innovative computational rhetorical analysis tools to investigate the use of citations in a corpus of academic articles. As a result of genre theory, our study uses graph-theoretic diagrams to extract and amplify expected patterns of repeated moves that are linked with stable academic writing genres. There is evidence to suggest that our computational strategy is as good as qualitative researchers who code by hand, such as Karatsolis and colleagues, in properly detecting and classifying citation movements (this issue). Pairwise comparisons of advisor and advisee texts reveal further applications for automated computational analysis as formative feedback in a mentoring scenario.

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

1. The Challenges of Learning Academic Genres That Are "Always Customized." 2.

Academic writing has a strange contradictory aspect that, when studied, explains why so many of us who aren't experienced writers have trouble finishing our projects, as pointed out in Pare's (2015) work on academic genres. Most importantly, the ability to create certain sorts of messages on a recurrent basis is available for certain types of texts. According to RGT, however, every occurrence of a certain academic genre has a "always custom" aspect, which undermines the concept that stable textual patterns will form as a result of recurrence. Pare. With Bakhtin mentioned, Pare shows that while the conditions under which some texts appear and recur may be generally stable, the precise conditions as well as the uniqueness of each writer enable or require that each text be distinct from the others. In the first paragraph, there is no indentation of the material.

By implying a certain distance between the genre, which is typically used in rhetorical genre theory to signify the social actions from which textual regularities emerge, and a genre instance, which is the single text that emerges from the distinctive social exigencies associated with a particular genre, RGT anticipates this "always custom" nature of genres. Learning and teaching are made more difficult by the fact that there is a wide range of variation within the same genre. Advisors and advisees also have challenges because of this. To be able to write in a new genre, you'll need more than simply grammatical accuracy. To be effective, it must be able to execute all of the genre-specific rhetorical manoeuvres flawlessly. Consequently, the meanings of the terms may alter significantly from time to time.

When it comes to RGT, maybe the most important lesson is that the repetitious texts we assign or analyse are simply the centre of much larger trajectories of stereotyped activity (A-91). When textual conventions are seen as a part of and in some cases a part of social relations by Pare's reading, this can be extended by pointing out that where there are repetitions at the textual level, these can be understood as signals shared by author and reader about the social activity (genre) with which they are cooperating. There is a wide range of information that is repeated, from whole chapters to only a few sentences. That being said, it's not just as simple as saying the phrases over and over again. A comparison of the same author's work in two different contexts may reveal that she has changed the wording but has produced a recognisable (albeit customised) example of the genre in both cases.

For this reason, it can be incredibly difficult for students to identify the "genre signals" that are connected with a certain type of writing in order to generate an example of a genre. It's possible that as instructors, we're making things more difficult for our students if we don't urge them to write about the kinds of situations and texts that are necessary in the larger discipline or community that they're hoping to join.

2. Analyzing Academic Writer Citation Patterns for Genre Signals

Using computational methodologies, we evaluate the possibility of creating a supportive environment for advisor-advisee mentoring in the area of academic writing and report our findings in this work. Pare's concern area is one that we aim to investigate and offer content that will benefit both learners and those assisting them in their endeavours. There are two key analytic passes that make up our approach: First, the texts in the corpus are treated as strings by text-processing algorithms, which means that they are no longer treated as individual words but rather as lists of words. Analytical methods such as word frequency and adjacency can be used when we approach the text in this manner. This is a huge step forward in the recognition of the importance of words in transmitting critical information.

We are seeking to find genre signals, which are often repeats of words or word combinations, that correlate to essential structural components, such as those that we may ask human raters to identify in a qualitative examination of text in the first pass. Repetitive social action (RGT) holds that generic expressions are fundamentally instances of repeated social action and that, in general, genre stability as indicated by regularised textual form arises from habitual responses to recurring social exigencies. During the first analytic pass, we are guided by RGT (Miller, 1994). As stated by Schryer (1993),

A more basic coding system than human raters would use is the result of our initial analytic run, yet it can sometimes provide results that are comparable to those produced by human raters themselves. Do you know how to go about accomplishing this? Often, it is because we uncover a consistent structure that might serve as a signal for a larger framework or a more subtle "rhetorical move" in the future (Swales, 1990). By recognising instances of propositional hedging, the endeavour to adjust claims to fit the weight of existing information that is so distinctive of scientific thinking, we can consistently define speech sections of various lengths as "science" or "not science" as we have shown (Swales, 2014). If a human reader would arrive to the same conclusion, it would not have done so by reading the same context signals or focused on similar details as a Hedge-o-Matic (Omizo& Hart-Davidson, in press). To put it another way, it's looking for a "key protein" in the complex molecular structure of scientific discourse. As long as the protein concentration is high enough, it can make a determination. Consistency in scientific discourse, not in the conventional textual sense, but rather in the manner Bakhtin (2010) conceptualises "speech genres" may be linked to Hedge-success. o's (see Bakhtin, 2010). Propositional hedging is essential to scientific discourse, and neglecting to use it is tantamount to breaking the social contract.

However, a second analytical step beyond the first text analysis is necessary for the Hedge-o-Matic to provide the results it produces. A graph is a collection of nodes that are connected by connections between nodes in this second stage of transformation. When we turn text into a graph, we open up a whole new set of analytical possibilities. There are more than just words on the graph now; there are also edges, which show the connections between those words. It is possible to understand not just the features of individual nodes, but also how they interact with each other and the overall structure of a network, using graph analysis tools. So, for example, it is possible to evaluate whether a particular phrase has a hedge move, but the Hedge-o-Matic also has the ability to identify hedges which are near a major statement in terms of the graph as well as its semantic content

The way we prepare the texts for analysis has a considerable influence on the results of both analytical passes. For these manoeuvres to work, interpretative work has to be done, guided by both the broad theories of genre and more particular ideas or hunches about the kinds of structures we seek to identify, isolate, or enhance. As an example of rhetorical structure theorization, the act of preparing texts for analysis can be viewed as a sort of pre-analysis. Instead of writing an exegesis from the beginning to the end, analytical "recipes" for discovering interesting rhetorical manoeuvres are being created. Consequently, we go into considerable detail about how we do our analysis here. It is not enough to just share our approaches with people who may find them beneficial; rather, we do so in the same way that we would in other sorts of rhetorical reasoning about texts.

Searching for Citation Moves in the First Analytic Pass

Using Karatsolis's qualitative coding method in this case with human raters to provide a computer analysis of the data is just too complex and intricate. As soon as possible, we had to devise a more clear plan. It was our goal to find out what kind of signals would be present in the corpus that could perhaps explain the perception of citation kinds by humans. Here, we'll explain in further detail how we came up with an alternative qualitative coding system.

Analytical techniques like those of our colleagues were developed to focus on more than one word or lexical item at once. Swales' (1990) idea of "rhetorical motions," commonly considered as significant patterns of speech within a specific genre or discourse community and performing crucial signalling roles, is what we were seeking for. Additionally, a text's rhetorical gestures reveal something about its position as an attempt by a given audience, in this case, readers of academic research papers, to respond to a recognisable type of urgency. A frequent example of this is the assertion of a claim, which is usually a new assertion of fact. To ensure that readers understand the author's unique perspectives and contributions, claims of truth are scrutinised carefully. ... Statements that explain evidence and hedges that qualify assertions based on the quality of the evidence are prevalent in science writing. We may see this in the form of textual regularity, which is a result of Miller (1984) and others' conception of genres as a sort of social activity that grows more stable through time and as genres become more stable.

Citation patterns are commonly blamed for part of this signalling activity, according to widespread consensus (Geisler, 1994; Prior, 2013). One of the most notable differences between our approach and that of our peers is that we are interested in developing assistive technology for both authors and text readers. Because of the speed and accuracy of these technologies, more junior members of a field may be able to learn from and collaborate with more senior members of the same subject.

Using the following questions as a guide, we began our data exploration:

What follows are some of the questions answered:

Classification patterns in citation moves that contribute to genre stability can we uncover evidence of? And can these patterns be accurately and consistently identified with the writing cohorts from which samples are derived (for example, experienced authors vs more inexperienced ones)? B.

b. Is it possible to create categories that will allow us to compare Karatsolis' findings with our own?

c. Is it possible to create analytic data that, if accurate, may benefit the advisor/advisee dyad for reasons such as academic mentorship or other comparable objectives?

Creating a Simplified Citation Analysis Coding Scheme

Since the beginning of citation analysis as a scientific field, there have been several proposals for qualitative citation classification schemes. Schemas that helped speed up search and retrieval might also be considered rhetorical in nature because of their intent to convince. By way of illustration, Moravcsik and Murugesan (1975, p. 3) have devised the following four-part taxonomy of in-text citations: (for an empirical application of the following categories see Cano, 1989). 1 Using six categories of in-text citation categorization, Chubin and Moitra (1975) propose a revised version of the Moravcsik and Murugesan (1975) typology in an attempt to measure the influence of a scientific paper on the growth of a discipline (in this case, high energy physics). The Chubin and Moitra (1975) method lays forth guidelines for how a research paper's claims and conclusions interact with the broader body of knowledge. Rules for academic writing may include the reference of relevant material that is crucial to the comprehension of the present piece and the denial of rival ideas in the subject of scholarship 2

There are two forms of in-text citation: communicative incentive systems and rhetorical ones. to acknowledge the intellectual property of the field in which they're employed, in-text citations can be used in this scenario (see Kaplan, 1965). The use of in-text citations as a rhetorical strategy to promote acceptance of the cited work is known as rhetorical citation. The value of a reference to the referring newspaper may be measured by categorising it into one of nine categories proposed by White and Wang (1997). A mix of in-text citation content description and/or a rhetorical manoeuvre can be used to describe this. For example, White and Wang (1997) defined the Analogy/Contrast/Comparison category, which suggests an act of interpretation in which the referenced authors compare and contrast their work with another. When a referring article uses data from another research source, the category of Data shows that the material was sourced from another source. 3 For the categorization of in-text citations to be successful, it must be handled as a supervised machine learning problem, which means it must be solved using a set of training data. Also discussed are 12 functions for labelling in-text citations based on the contrastive use of references by the referring publication and the positive or neutral valence of the reference's use by the referring work (p. 104-105). Textual components, such as parts of speech and metadiscursive phrases, were incorporated into their training data as well as their classification algorithms in order to enhance their classification methods.

The results of our investigation led us to classify in-text citations into four distinct categories: extractions, groupings, author(s) as actors, and non-citations. Extractions: A large part is due to our initial exploration of data from the Springer Open Journal archive, which is constrained by Harvard-style citation rules and Springer Open mark-up templates used in the presentation of research articles, and which we discovered during our initial exploration of data from the Springer Open Journal archive. Springer Open Access journals were screened for 505 research papers to create our categorization method. All of these publications are peer-reviewed and adhere to academic genre standards, such as the IMRaD format, which is often used to structure scientific and social science journals (see Christensen and Kawakami, 2009; Hannick and Flanigan, 2013; Salager-Meyer, 1994). The article's meta-data (authors' names, publication dates, affiliations, and document object index) were scraped from the original source document, as were the article's entire text without pictures and the works cited list. Using the Springer OpenAccess filtering tool, only articles classified as "Research Article" were taken into account for this exploratory study.

This benchmarking project has two objectives in mind. We ran a series of experiments to see if our categories could account for every potential sentence type that may be found in an academic research publication without the requirement for subject expertise. A corpus of 505 research articles from Spring OpenAccess served as the basis for this study. By using this corpus, we could optimise the syntactic and argumentative signals that each sentence in a research article could provide for computation, while also reducing the likelihood of overlap or ambiguity between categories and the requirement for domain-specific knowledge of sampled journals. Moravcsik and Murugesan (1975) show how domain knowledge may be employed in a coding situation involving the type 3 category, development of juxtaposition. While a non-expert might be able to tell that "Our study builds upon the work X" or "Our study varies from X in the following respects" without metadiscursive clues, a decision between evolutionary and juxtaposition may be more difficult in the absence of metadiscursive signals." Because Moravcsik and Murugesan (1975) and Cano (1998) both use type 1 perfunctory citations, it's important to have a solid grasp of the subject area before making such use of the citations consistently (1989). Because the inclusion of a reference might be interpreted as rewarding a colleague while still being considered as rhetorically fitting to the argument, the reward system proposed by Cozzens (1989) may induce conceptual ambiguity This research and most other computer techniques lack an interview, which is necessary to hypothesise about the writer's intents because of the idea of reward. Most other computational approaches also need us to move above the lexical level (similar concerns are raised in Teufel et al., 2006).

Second, we were able to create a sentence parser that automatically distinguished between citational and non-citational sentences and then sorted these phrases into more specific categories based on lexical information using the Springer Open Access corpus. We didn't utilise our sentence parser to code the sentences from Karatsolis' data, but we expect that it will be useful in the future for other applications that need processing of bigger volumes of text.

The "shallow analyzer" method provided by Marcu et al. motivated us to build a lexically-driven yet rhetorically informed coding system (1997). Pre-marked cue phrases with specific grammatical functions and rhetorical applications were gathered by Marcu (1997) in an effort to scan natural language texts for their rhetorical purpose. Cue phrase "Although," for example, can be used to break up the text, but it can also be used to make concessions because of the limited use of it in English. To make an argument more compelling, it is common to use a trigger phrase like "yet" (Marcu 1997 p. 101). (Marcu, 1997, p. 101; Marcu, 1997) Our coding approach considers all four citation types to be relevant in their lexical differences, even when phrases like "despite the fact that" and "despite the fact that" help to delimit the alternative rhetorical readings of a sentence. Even in the lack of further context or intertextual information about the author, the presence or absence of an author's name in a phrase is regarded to have significance within the sentence itself. With our coding system having limits, we feel that classifying citations into four categories can give important information on academic writing citation practises, as well as a horizon line between material from mentors and those of their protégés. First, we'll go through the requirements and rhetorical contributions to the subject for category 1.

4.1 Extractions

Excerpts include in-text citations that provide a concept from another source without explicitly referencing that source. A parenthetical reference is a frequent example of a paraphrased idea from a source that is attributed to the source. Kemner et al. (2015) claim that stress and diathesis interactions are prevalent explanations for psychopathology's emergence (Monroe and Simons [1991]). Using conventional citational parentheticals to emphasise the material from Monroe and Simons' piece is no longer the preferred method of underlining their work as active agents in the study of contemporary research. Motives for making this shift include wanting to build one's own voice as a specialist in a certain topic or developing one's own personal style (e.g. a choice to describe conclusions as facts rather than narrating research as a process). According to Swales (1981) and Swales (1982), the reference of Extraction falls within the category of "non-integral citations" (2014). There are two ways to credit a referenced author in these citations: in parentheses or as a number index.

4.2 Groupings

An example of this type of grouping is the phrase "in-text citation phrases that include three or more sources within a parenthesis or brackets" The following from Kemmer et al. (2015) serves as an example:

Of particular note, several studies have found that having previously experienced depression or another mood illness to be an important risk factor for future recurrences of those conditions (Judd et al. [2008]; Keller et al. [1983]; Perlis et al. [2006]). The in-text reference from Kemner, et al. (2015) is categorised as Grouping in this case study because it refers to three sources that have uncovered evidence that significant life events/difficulties affect the emergence of unipolar depression and bipolar diseases. "Parenthetical plonking," or "nods all around to past scholars," is what Swales (2014) refers to as "Grouping," a rhetorical strategy he employs (p. 135). On the other hand, "plonking" or "grouping" strategies are viewed as attempts to synthesise a range of individual findings that are consistent around a larger issue. We feel that this method, used by researchers such as Kemner et al. (2015), acts as a condensed literature review since it allows them to exhibit their knowledge of important sources while also showing their connection to a broader network of academics studying mood disorders. There is another example that may assist to understand their thesis more clearly provided by Kemner and his colleagues (2015):

Also found in earlier research by Bender and Alloy (2011; Hunt, et al., 1992) and Kessing et al. (1998), however it was proposed that these occurrences may be linked to the syndrome itself (Bender and Alloy [2011]; Kessing et al. [98], [2004]). (Hunt et al. [1992]; Kessing et al. [1998], [2004]). Researchers have found that (Kessing et al. [2004]).

The authors of this specific study are looking at the history of studies that are comparable to their own and that both confirm and differ from their own findings. Leighton (2014), for example, utilised what we term a Grouping in-text citation to gloss whole fields of research through parallel references in the Journal of Evolutionary Education and Outreach.

Since the dawn of time, scientists in all fields have been concerned with the protection of public assets (Hardin [1998]; Boyd et al [2003]; Bowles [2006]).

As previously indicated, the parenthetical list is supposed to reflect an overall theme or concern with each reference in the list. This list is used to find current research in relation to a certain group of academics within the field of study, which is how the Grouping categories obtain their grammatical grouping of sources in a parenthetical list. Swales and Najjar's "construct a research space" (CARs) strategy might support such a shift (1987). For producing research article openers, Swales and Najjar (1987) established a four-step process. A research article's introduction typically includes four sections: (Move 1) signalling to the reader the importance of this research; (Move 2) summarising relevant prior work in the field; (Move 3) indicating an insufficiency in prior work; and (4) explaining how the present study will address current gaps of knowledge (1987, p 178- 179). Swales and Najjar's moves 2 and 3 may be considered a way of "creating research space." in our study's Grouping category by conveying information to the reader in one sentence.

Extractions and groupings are both considered non-integral citations in Swales' (1990) definition since their basic ideas overlap considerably. Extinct ideas can be identified in the author's argument by looking for proof in a parenthetical line that includes a summation of study and several references. The ratio of summarization in the argument is one of the distinctions in rhetorical value. A single line from one or two sources indicates that the source has been less filtered or homogenised, emphasising what particular agents are saying in relation to the present piece's theme. More than one source is needed to make a single statement on the present article's subject matter, whereas a single source is needed to make a statement about a single subject matter. Another approach to distinguish between extraction and grouping is to see them as actions of enrolment. Every citation is, in the end, a way to collect data from a variety of sources. When evaluating enrolling for our purposes, we look at how prominent it is in the context of the phrase. Three or more sources included in parenthetical notes demonstrates a wide breadth of interest in a given subject and shows the present writers' current effort in learning about other voices in the area. If you use an Extractions in-text citation instead, you'll have more room to draw attention to your engagement in the source material.

The Author as an Actor (or Authors) (s)

To use the term "actor" in an in-text citation, the author of a piece of cited research must either be the subject or object of the sentence in which they are cited, the subject or object of subordination in which they appear, the source of a direct quotation, or the subject of a possessive contraction linking the author's name to a piece of cited research. In-Text Citations Author(s) and Actant(s) A date of publication reference in parenthesis or a bracketed number reference in brackets must accompany these in-text citations. This citation style, which Swales coined the term for, might be seen as a more condensed form of the "integral citation" style, which was originally characterised as (1990). This syntactic pattern is used by Thompson and Ye (1991) in their research of so-called "canonical" citational forms, which contains the proper name of the author followed by a year in parentheses or brackets, which functions as either the subject or the object of the phrase.

No date of publication or pronoun attributions are taken into account in the Authors as Actor category. "They further contend that," for example, is not recognised as an Author(s) as Actant(s) citation by this coding approach; instead, the sentence pattern is classed as a non-citation. "I" and "us" pronouns are used in extractions, which appear to be full citations, although they are not. "(see Smith, 2007)," would not be categorised as an Author(s) as an Actant(s) citation type since the source is subordinated inside the sentence by a parenthesis. On the other hand, the "see" author suggestion is incorporated into the text's main argument and is classed as an Author(s) as Actant phrase (s).

This example from Keown-Stoneman et al. (2015) shows a circumstance in which the author of a work that is being referenced becomes the subject of the sentence: ""

Among this high-risk group, Duffy et al. ([2010]) hypothesised that some psychopathological signs may be precursors to the onset of bipolar disorder.

"Duffy et al." is the subject of a sentence when it refers to a theory concerning the likely causes of bipolar disorder that they "proposed" In this perspective, Keown-Stoneman et al. (2015)'s research space includes a diegetic dimension in which the referenced authors act or are acted upon at the sentence level.

In Keown-Stoneman et al. (2015), an example of an Author as Actors in-text citation can be found, where the referenced author is the subject of the predicate.

Detailed information on the collecting processes is available in Duffy and colleagues ([2007]).

Authors Keown-Stoneman et al. (2015), for example, indicate that more study has been written about by Duffy et al., meaning that Duffy et al. has carried out their own investigation. For example, "Duffy et al. ([2007]) provide a more extensive description of these procedures," becomes "Duffy et al. ([2007]) offer a more detailed description of these approaches." in the passive voice citation above.

Example: A statement from Correa-Bahnsen-et al. (2015), wherein each author is a subordinate clause, provides as an instance of in-text citation in which the named authors appear as subordinate clause objects:

Afterwards, we looked at the results of a cost-sensitive logistic regression (CSLR), which was calculated according to the literature's default parameters (Correa Bahnsen et al [2014a]).

When Correa Bahnsen, et al. (2015) allude to an article in which they provided methodologies for a cost-sensitive logistic regression, they are referring to a previous paper in which Correa Bahnsen, et al. were successful in their endeavours. As a result of journal markup, the parenthesis are disregarded in this example, and the listed authors are assumed to be the object of preposition "in" in the following phrase.

According to Thompson and Ye, the use of an author's name in conjunction with a reporting verb like "display," "confirm," or "give" is an act of interpretation (1991). The writer's attitude toward the given information is shown through these actions of interpretation. The orientation may be used as a reward mechanism to recognise the achievements of the cited author, or it may be used as a punishment mechanism if there is agreement or disagreement between opinions or facts (Charles, 2006 p. 322). According to Paul (2000), naming the author(s) of a reference is a reflection of the source's significance in the field. There are two options, according to Paul, if there is no number or parenthetical reference: (p. 199). Authors' names serve as a convenient cue for readers to focus their attention on a certain passage. Their works, Thompson and Ye (1991), Paul (2000) and Charles (2006) are all built around a core of assessment that serves as the basis for their nomination instances. There are a lot more textual features involved in these evaluations than our coding technique can handle. It is important to analyse the type of reporting verb employed in a favourable, negative, or neutral assessment (Thompson and Ye, 1991 p. 372). The polarised valences of "oppose" and "agree" would be inverted in the citations "We oppose Smith (2000)" and "We agree with Smith's (2001) findings." Our coding method, on the other hand, only takes into account how a verb functions in relation to a given author, rather than the type of verb. Do you think our curiosity was sparked because we were introduced to specific practitioners in the area by referring to an author as a topic or object of an activity? Authors can then use this rhetorical move in both a positive and a negative way, thereby affirming or extending related work while also complicating or challenging it. This results in an entirely different kind of engagement with the source material than is possible with the Extraction or Grouping citation types.

4.4 Non-citations

This category is for sentences that don't fit into the Extraction, Grouping, or Authors As Actants citations. As an example, attributions that use a pronoun in place of the named agent and do not contain a parenthetical date or a parenthetical with the name and date or a bracketed numeric reference fall under this category.

Following our coding method, it is logical to assume that typical attribution activity, especially if it is an extension or enlargement of an earlier attribution, is classified as non-citations by our categories. Due to limitations in the shallow parsing tool and an effort to eliminate subjective coding decisions, this elision was made. Citational incursion, in which writers are signalling their compliance with research standards and signalling additional arguments, is the sort of behaviour we are seeking to capture using shallow parsing and network graphs. For example, a nod, a rounded sweep of the hand, or a pointing motion are all examples of rhetorical gestures in the style of Gilbert Austin or Francoise DelSarte, respectively.

Analytic Pass 2: Seeing Citations in Graphs as Essential Structures

Using citation coding scheme decisions, we have constructed a network graph structure that uses the numerical codes 0-no citation, 1-extraction, 2-grouping and 3-author as a node list of [0, 1, 2, 3]. As a result, the graph of the network consists just of these four nodes. To distinguish one node from the one directly adjacent to it, lines are placed between them. The original order of the source text dictates the node sequence. Here's an example from Lemieux (2015), which has been labelled as follows in our SpringerOpen investigation:

Digital images that include spatial information are commonly known as geotagged shots in the digital photography business. There are a number of techniques to geotag a photo, some of which are manual and others of which are automated. (1) Digital cameras with GPS built-in or connected to the camera allow for automatic geotagging. Some camera manufacturers, such as Casio, Olympus, and Nikon (Valli and Hannay [2010]), also provide devices that have a built-in GPS receiver. 1)

[(0-1), (1-0), (0-1)] is how the edge connections between the phrases would look with citation tags. Since all the tags point in the same direction, we may think of the previous set of edges as a path that links 0 and 1 and 0. The likelihood is that each graph will have nodes that link to themselves (or to other nodes in the same class) in order to avoid establishing differences between various phrases and to represent each sentence as a class (e.g. 0-0). Graph representation creates the idea that nodes with the same class assignment are identical, despite the fact that this is not the case. Many edges are created between nodes due to this oversight, resulting in an arrangement that is almost identical to the annotations in the original text. Self-looping links the nodes in this network structure, giving it both direction and a plethora of edges. Although there are self-loops and repeated edges that appear to be present, the graph structure is an Eulerian path, in which each vertex has an equal number of incoming and outgoing connections, except for the vertex at either end of the path. This graph structure is equivalent to a network graph in which each node has a class value and an individual identification number. It would be possible to tell one node and edge apart from the rest in this circumstance. Given the importance we placed on preserving sentence order as it appears in natural language texts (in English), we chose an Eulerian route network graph structure as the basis for our data model. We can keep track of the text's citational and non-citational motions by modelling it as an Eulerian route. Using an Eulerian route, we will be able to extract elements that authors use by default, such as dividing material into parts with a beginning, middle, and end, among other things. As a result of the text-to-graph structure, a computational operation may be applied to an array of coding decisions. Network structures may be studied using these approaches to extract attributes that can be used to develop a model that compares advisor and advisee texts. Text-to-graph structural elements are explained in the next section, along with why this feature was used in the context of a rhetorical research on how to include citations into the text.

Indicators Based on Graph Data

According to the network adjacency matrix, the citation nodes' eigenvalues

This feature includes the eigenvalues of the graph's three different sorts of citation nodes (i.e., 1, 2, 3). If you want to know the number of nodes in your graph, you need to take the greatest eigenvector value and multiply it by that number. Known as the eigenvalue, this highest value of an eigenvector serves as a measure of the matrix's fundamental properties. We utilise these eigenvalues as a baseline for comparison in our study of adviser and advisee texts since they pertain to significant structural elements of the graph. Because they contain self-loops, asymmetric adjacency matrices are common in our graph architectures. Consequently, adjacency matrices will typically contain values that are unduly biassed toward edges between non-citational nodes (0-0 or sentences without in-text citations that follow other sentences without in-text citations).

We only save the eigenvalues of the adjacency matrix that govern citational nodes when graphing advisor and advisee texts since the citational practises of advisors and advisees are our major focus (1, 2, 3). It's no secret that advisor and advisee writings rely heavily on non-citational phrases. This is due to the general character of academic writing. A high Eigenvalues or Eigenvalues with substantial effect on the nature of the network isn't surprising because the 0-node vector has high Eigenvalues or Eigenvalues. That is why our focus is on how to utilise citations strategically rather than how to use them liberally across a whole text.

It is possible for a subgraph's size to vary.

With this feature, nodes with an inbound or outbound link to 1, 2, or 3 other nodes are removed from the graph altogether. At the spots where deletions have occurred, the Eulerian trace has been split. To reassemble the Eulerian trail with the following path, for example, it would be necessary to eliminate all of the 1-citational edges. This would result in three separate sequences or subgraphs: 0-0-0-0-0; 0-0-0-0; 0-0-2. This characteristic is similar to that of "network robustness" (Albert and colleagues, 2000), which is a measure of how well a network can withstand node deletion while still preserving its connectivity. Although Albert et al(2000study)'s on network robustness is applicable to the sorts of graph topologies developed in this study, it is not immediately transferrable. It is of particular importance to compare the responses of exponential and scale-free networks to attacks and node malfunctions, respectively, as Albert et al. (2000) have shown in their study. Scale-free networks have a network architecture characterised by a limited number of essential nodes with many edges and a large prevalence of nodes with a smaller number of edges. A powerlaw distribution is only marginally supported when the nodes of our graph structure are typical of classes 0, 1, 2, and 3 in the Springer Open Access corpus and advisor and advisee texts. Similar arguments may be stated in favour of using the lognormal distribution over a powerlaw distribution. As a result, the Eulerian pathways we used in our study lack the same robustness as those found by Albert and colleagues (2000; see also Albert, et al., 2000).

Second, Albert et al (2000) investigates websites on the World Wide Web, rather than typical webpages, as nodes in a vast directed network. There are no citational and non-citational sentences inside a single article, but the sequence in which the words are presented is what makes the difference. Espan.com could be linked by a group of regional sports websites, but these sites are free to develop additional relationships with other websites and benefit from fresh incoming links. Because nodes in our network topologies reflect separate sentences in a text, the order in which they appear is predefined. A scale-free or exponential network's connection may suffer if a node is deleted at random, but data from a deleted node may still reach another via alternate channels, raising doubts about the applicability of Albert and colleagues' (2000) concept of network resilience. In the Eulerian routes we are using as models for citational practises, every single node or edge deletion will result in a short in the network. As a matter of fact, on an Eulerian route, because each node has only one incoming and one outgoing edge, each node delivers the same amount of connectivity (with the exception of the very first and very last nodes in the path). Eliminating any node or edge will prevent data from travelling through the network in a way comparable to but at a different place than it would otherwise because of the lack of redundancy and other paths. As a result, our graphical representations of networks do not match the resilience or disruption concepts for scale-free or lognormally distributed networks.

The size range of the subgraphs created by the deletion of edges that contain a citational node and the number of edges that are eliminated are used to assess the robustness of the method. For each subgraph, the size is divided into quartiles, with the first and third quartiles being used to measure the subgraph's size. A subgraph fragment's usual size is used here as a way to estimate the degree of disruption caused by the deletion of citational edges or the relative size of the Eulerian network components that citational edges help to link. We use the interquartile range of sizes instead of the mean of sizes to adjust for the overall skewness of subgraph sizes. This indicates that citational types are less common and more broadly dispersed in a text's subgraphs with a bigger interquartile size range than smaller interquartile size ranges. There are smaller pieces after removing the edges of citational types, which indicates that the kind of citation is more common in the text and has closer connections to other citational types.

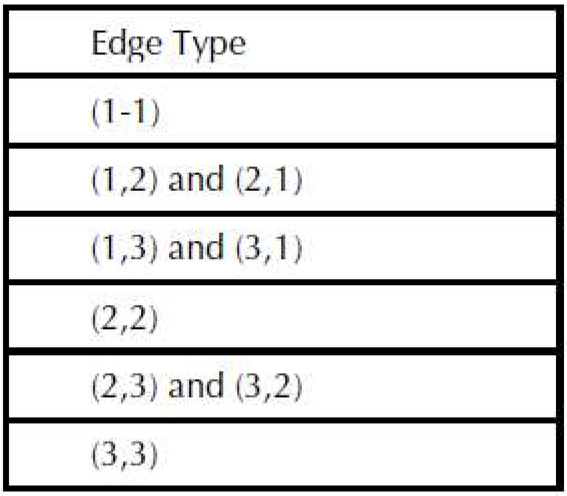
6.3 A thorough examination of the distribution and placement of citational edges is provided.

This feature's objective is to account for the network graph's citational edge positions. The network graph sequence is broken down into the following percentile ranges in this manner. For example, in a network series or research article, 0-30 represents the beginning; 30-75 represents the middle; and 75-100 represents the conclusion of the sequence or research article. Percentile ranges of citations are recorded for each kind and then divided by the total number of citations in the section to get the relative frequency per section for that section. According to Cano's research, which was repeated using this feature extraction approach, citations are an effective research tool (1989). Even though research articles allow authors to insert citations at any point in the text (and frequently do), research by Cano (1989), Voos and Dagaev (1976), and Ding, et al. (2013) (see also Tang and Safer, 2008) has found that citations in scientific research articles are primarily located early in the document—in the Introduction, Literature Review, and Methods sections. When Hu, Chen, and Liu (2013) studied the Journal of Infometrics' 2013 corpus, they found that more than half of the in-text citations occurred in the first 30 percent of each article, based on their findings (p. 891). According to Paul's research on the literature on chaos theory, half of the citations come in the article's introduction section, which is also where the majority of the research is done. 6 Using 505 research publications from the SpringerOpen database as a benchmark, a similar concentration of citations was found. We found that 46.8% of citational edges appeared in the first 30% of our network sequence, which is substantial. 7 Consequently, this feature serves as an indicator of how well a writer adheres to the genre's citational practises. As a general rule, a larger concentration of citations should be seen near the beginning of the network. On the other side, the absence of such a density might indicate a departure from the genre standard. Furthermore, the areas where citations for Extraction, Grouping, or Author(s) as Actant(s) kinds cluster within a text may offer further information about the precise general limitations put on each type of citation.

6.4 Edge reciprocity

In a loop of edges, two vertices in a directed graph exhibit reciprocity if and only if: There are nodes in the network that point back to each other, and so on (Newman, 2010 p. 204).

Each node in our Eulerian route represents a different sentence in the original text, therefore they are all distinct from one another. When we talk about reciprocity, we're talking to node classes that are related to one another in the graph's linear arrangement. A reciprocal arrangement for nodes in classes 0 and 1 would be the arrangement of their nodes in the order of 0-1-0. It is possible to represent the degree of edge reciprocity as a percentage of the total number of edges. Non-citational node classes that link to one another in the network produce the same kinds of edges that citational node classes did. Due to the huge number of non-citational nodes connecting to one other, this is the case. As a result, the edge type 0-0 is eliminated. Each of the remaining edge reciprocity permutations may now be represented by a single value thanks to this technique (see Figure 1).



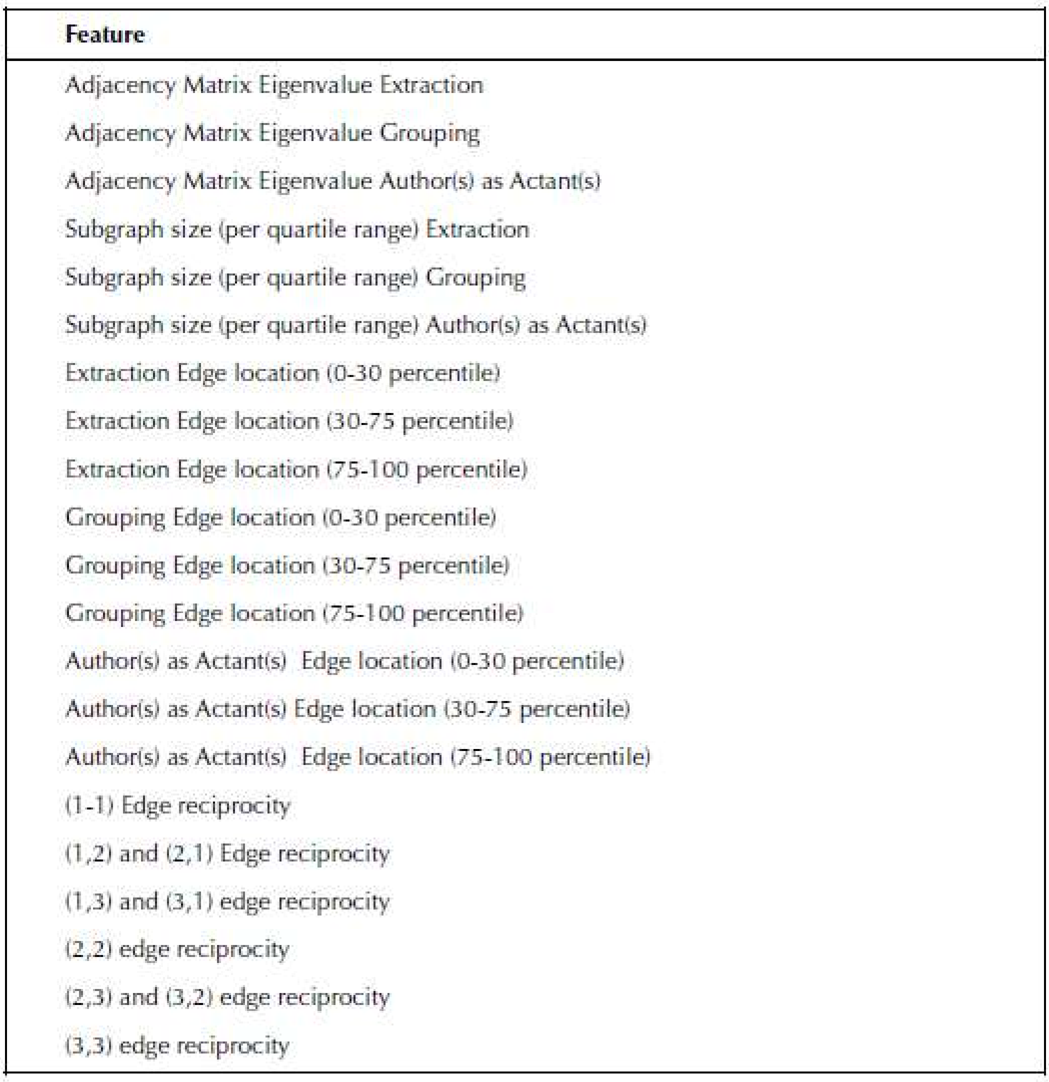
Edge Type
(1-1)
(1,2) and (2,1)
(1,3) and (3,1)
(2,2)
(2,3) and (3,2)
(3,3)

Figure 1: Edge Reciprocity Types.

Author(s) and Actor(s) as Actor(s) are cited in the text in such a way that they tend to inform one another, and this feature may be employed for this purpose. Conventional discourse would lead us to believe that there is more reciprocity between the various sorts of extraction and grouping than there actually is. Both the Extraction and the Grouping citation types would be utilised to extract concepts from the source they are citing in this scenario. Extraction and Grouping citations are more completely incorporated into the argument's style and construction, and are supplied as pure information rather as citations, as Voloshinov points out in his essay. It would be more accurate to classify author (or author(s)) quotations as "reported" rather than "quasi-direct" speech, because they place an emphasis on the "what" rather than the "what" and the "how" of what is being said or done (Voloshinov 1973, p. 117). (Voloshinov 1973, p. 119). By using terms coined by Latour and Woolgar (1976), the distinction between "extraction," "grouping," and "author(s) as actant(s)" can better be described as one of "facticity," where parenthetical references draw less attention to other research efforts than the clausal incorporation of an author's name, making the sentence appear more factual on the surface (p. 80-81). Extraction or Grouping moves may be more convenient because the goal of a reference is to give uncontroversial information or acknowledge that comparable work exists, omitting the need to offer context or qualification both philosophically and aesthetically, as stated by Cozzens (1989). (see Cozzens 1989, p. 443). To back this up, we looked at the SpringerOpen database, which comprises over 500 peer-reviewed papers. Nodes engaged in extraction had the highest mean edge reciprocity scores (1-1). Mean Edge Reciprocity Score for Extraction and Grouping is second only to Extraction and Grouping in terms of average score (1-2, 2-1). Following closely behind is the link between Extraction and Authors acting as Actants, which has the third highest mean edge reciprocity score (1-3, 3-1).

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

Figure 2: Graph Feature Array.



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

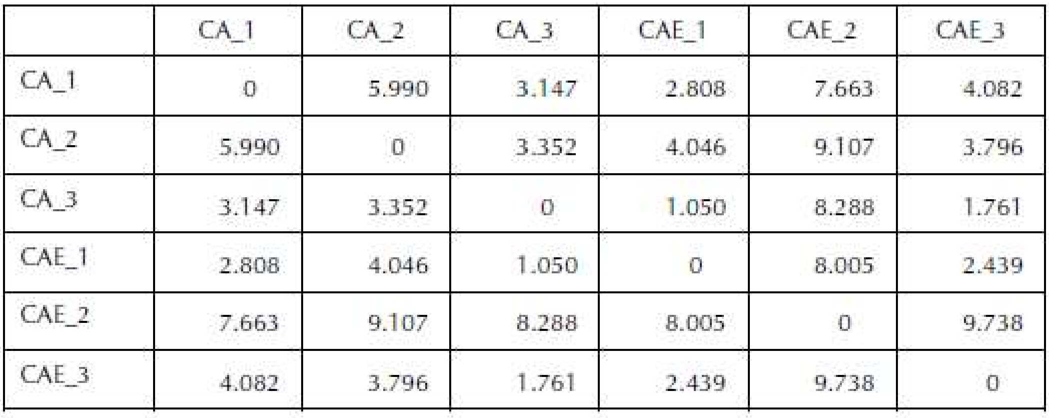
The graph feature values tabulated in the array are then used to calculate the Euclidean distance between network graphs.

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

Our two analytic passes will be discussed in detail before we return to our three framing questions. This research is still in its infancy, and more work need to be done to determine whether or not the findings can be relied upon. To begin, we should point out that our corpus was rather small by big data or text analysis standards. In addition, the writings do not adequately depict the phenomena researched. We chose to call this study exploratory rather than confirmatory or even descriptive for the following reasons:

The results we've seen thus far are encouraging, and we feel that greater investigation into this topic is needed. The citational patterns observed by Karatsolis are matched by our findings of genre signals (this issue). It will be necessary in the future to conduct more extensive investigations to draw more reliable results. Our exploratory study ideas, on the other hand, yielded fruitful outcomes.

Are there classifiable patterns in citation changes that contribute to genre stability and possibly explain the similarities and differences in the writing cohorts from which the samples are collected (for example, more experienced writers vs. less experienced authors)? If so, what are they? The first question looks to have an affirmative response. For this study, we looked at articles from the Karatsolis dataset from two different groups of advisers and advisees. The cohorts include chemistry advisor and advisee textbooks, as well as materials science advisor and advisee texts. Using Karatsolis' original comments and our own developed coding system from the first analytic run, we manually identified citation patterns. Figures 3 and 4 show the distance matrices used to illustrate the network graphs in this section.



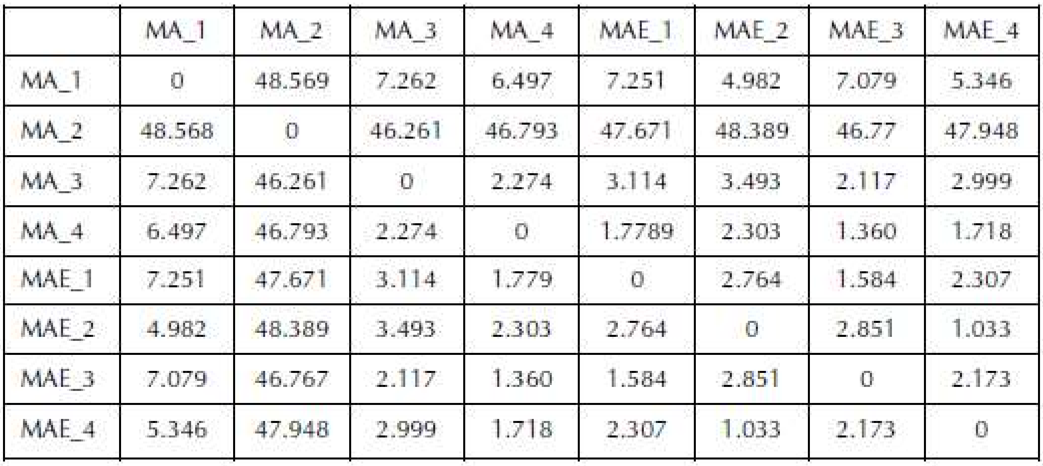
	CA_1	CA_2	CA_3	CAE_1	CAE_2	CAE_3
CA_1	0	5.990	3.147	2.808	7.663	4.082
CA_2	5.990	0	3.352	4.046	9.107	3.796
CA_3	3.147	3.352	0	1.050	8.288	1.761
CAE_1	2.808	4.046	1.050	0	8.005	2.439
CAE_2	7.663	9.107	8.288	8.005	0	9.738
CAE_3	4.082	3.796	1.761	2.439	9.738	0

Figure 3: Pairwise similarity scores for chemistry cohort.

Three textbooks, CA 1, CA 2, and CA 3, are written by one professor who leads the chemical group. In the chemical cohort, one advisee has written three texts: CAEs 1–2 and 3–4. The distances between the samples are shown in Figure 3 (as shown).

the following texts are written by an advisor in the material science cohort: A, B, C, D, E, and F One, two, three, four, five, six, seven, eight, nine, ten, eleven, twelve, thirteen, fourteen Data from the material science cohort is depicted on Figure 4.

Using this method, we feel we have achieved what we set out to do: provide an easy way to compare texts between and within cohorts. What do these apparent differences show, then?



	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?

According on the facts, the answer to this question appears to be a little more qualified yes. To get a sense of where we are, we may look at the average similarity ratings of the writings written by different writers. The chemistry advisor's texts (CA 1, CA 2, and CA 3) have an average pairwise distance of 4.163 points on the chemistry advisor's text distance scale. CAE 1, CAE 2, and CAE 3 texts have an average pairwise distance of 6.72700061 characters (in hexadecimal notation). The material science advisor's texts (MA 1, MA 2, MA 3, and MA 4) had an average pairwise distance of 26.276. In general, the material science advisor's writings are separated by an average of 2.119 points in pairwise distance. Comparing texts by different authors, it appears that the average pairwise distances follow certain predictable patterns. We would predict less variance in citation patterns in the more novice advisee texts compared to the more experienced advisor texts because (1) advisees are less experienced authors than advisors, and (2) less experienced writers frequently have less opportunity to perform in a range of genres. Greater experienced writers have more freedom in terms of their writing styles than less experienced authors (see Berkenkotter and Huckin, 1995, p. 117-144). As a result, we may predict that the citation habits of advisees will become more regular over time. The chemistry cohort advisee exemplifies this. Extraction citations are often used by the chemical adviser in the texts sampled. Extraction citations are also heavily emphasised in the first third of articles in chemical advisee manuals. Author(s) as Actant(s) citation is commonly used towards the conclusion of a document, particularly in the final third of the publication, while presenting the results of a research or experiment. It is consistent with the IMaRD framework of a scientific research report, and it is also consistent with the chemical advisee's efforts. Consequently, the article's technical content, which occurs early on in the paper, is meant to be more descriptive in nature. Because of this, it is more common for citations to be excerpts meant to set precedence for an approach quickly. There should be a greater emphasis on previous research in the final discussion section in order to properly interpret the findings and emphasise their relevance in relation to other studies in the area.

As can be seen from the high frequency of non-citation moves and the prominence given to citational moves in the first third of articles, there is even greater uniformity across members of this cohort. Using citational motions to build a research space (in the Swales sense) and then adding new work to this rhetorical stance is common, according to the material science adviser.

Advisor texts will also be more varied because of their higher status since advisers have (1) more disciplinary writing skills; (2) better access to a wider range of literary forms; and(3) more freedom to break from convention. In a dramatic way, this may be seen in the texts for material science advisors. The adviser and advisee textbooks in material science do not have anything like the MA 2 text. In our coding method, MA 2's text suggests a total of four citations, which is a low quantity. As in MAE 4, the total number of citational movements in MAE 2 is 9 (MAE 4 = 32 vs. MA 2 = 54), which is similar to the number in MAE 4. Only the first quarter of the series has the MA 2 citational changes, which is noteworthy. A higher number of these citational moves are spread out among the different texts in MA 1, MA 2, and MA 4, suggesting that these texts are more heavily dependent on these moves to build an argument and illustrate conclusions than other texts. It's easy to see in the MA 2's text that the focus is on explaining the findings of an experiment on nanoparticle motion. Our macro-evaluation is bolstered by this. The first sentence of the text has a Grouping citation, which shows this:

There are a broad variety of unexpected behaviours that may be seen when it comes to nanoparticles made from inorganic materials (1-3) It is the goal of MA 2's research to at least theoretically confirm traditional knowledge of nanoparticles. In this case, it is reasonable to expect that less citational labour will be required. While MAE 4 aims to challenge the status quo, as stated in the following goal statement:

We provide here a method for producing parallel carbon fibres and extended, ordered networks of multiwalled nanotubes that may be utilised to form layered multiple connections using a modified CVD methodology built from our carbon nanotube growth strategy.

Taken along with MA 2 text's extreme pairwise distance figure, we can observe that the material science advisor and advisee text samples are quite near. Figure 4 shows that the MAE 1 and MA 4 texts are quite close in Euclidean distance.

The contrasts that Karatsolis points out between advisors and advisees in the texts, on the other hand, do give some useful targets for differentiating between the two. One of the most notable variations is the way in which the sources are evaluated. Citations that include references can be found here. The first was submitted by a chemical advisee (CAE 2), while the second was submitted by a chemical advisor (CA 2). Keep an eye out for the advisee's use of evaluative language in her sentence:

Advise: Long and Norrish31 (-136 + 25 kcal mol-1) got a measurement that is more in line with the computation than any of the others reported by Cox and Pilcher31.

Advisor

The proportional-derivative and minimal variance adaptive controls were employed by Meline and colleagues (3) to get around the learning periods of adaptive controllers.

Aside from the fact that the advisor's use of quotation marks is more descriptive, he also describes the authors' work as "close to issue resolution" rather than "better" or "more accurate". Even if both texts follow the same pattern of elaboration, words like "really" and "better" jump out in the advisee's text because they are plausible indications that are typically lacking in the advisor's writing.

This is one of just a few changes that can be clearly seen at the sentence level. After coding and comparing hundreds of sentences across a large number of texts, the majority of the differences become apparent (as demonstrated by the distances shown above). To put it another way: When two writings are similar, it is not merely because they utilise the same set of vocabulary to describe the same subject matter. Instead, the entire texts are constructed in a way similar to the snippets at the macro level.

An analysis of macro structure at this level may help to discover areas where a writer is producing a work that is significantly different from other writers. The text produced by the chemical advisee is markedly different from the texts written by all of the other authors, both advisors and advisees, in the similarity scores table above. Accordingly, we might guess that the writer is unfamiliar with some part of citation practises in their field of study.

Due to Berkenkotter and Huckin's (1995) findings, writers commonly utilise citational procedures to highlight the newness of their findings in their scientific writings. The macro-perspective offered by the graph analytic may assist reduce the breadth of research subjects to a more defined range of options. Before moving on to the rest of the texts in a cohort, researchers can identify the outlier text, analyse the text's important parts, and gain a broad idea of what deviations to look for before further investigation.

The most important feature of our strategy is not that we can exactly reproduce each human-coded judgement using a methodology like Karatsolis', but rather that we can. Advisers and their advisees may benefit from focusing on a specific set of rhetorical techniques (e.g. citations and their utilisation) to better improve awareness of the need for more targeted readings, comments and revisions. Automatic analysis can help teachers and tutors better grasp the rhetorical strategies that beginning writers might apply and whether or not these strategies are reflective of the types of strategies employed in a specific field.

When it comes to helping students with their academic writing or making professional selections, can we come up with analytical results that if they are correct, would be beneficial to the advisor/advisee dyad?

We'll have to wait and see, but we're cautiously hopeful based on what we've seen thus far. With Karatsolis' observation that advisees supply substantially more detail than advisers, we may see opportunities to detect them and highlight them during the writing and editing processes.

One of the most intriguing possibilities we see is a connection between citation categories and Karatsolis' findings. If, for example, we wanted to uncover patterns of more or less elaboration around citations, we may categorise citations according to the author as the actant category (i.e. author as actant moves mean there are more likely to be elaboration). This is an example from one of the chemical engineering texts that the advisee is studying:

Researchers Felinger and Guiochon [20] showed that optimising the experimental settings in displacement systems may be accomplished using a simplex technique that had been slightly adjusted. Their results, however, are confined to materials in the stationary phase with particle sizes ranging from 5 to 20 microns since they utilised the equilibrium-dispersive model.

In the face of such a drastic shift, there are two compelling reasons to expect further explanation from the author. Furthermore, the selection of an author as the active citation type is more time consuming (compared to the selection of an Extraction or non-integral citation type), and it indicates that the researcher wishes to engage more directly with the specific works of other researchers, as Paul (2000) argues (p. 2000). Because of this, it is reasonable to expect that the author's capacity to compose a whole sentence will frequently be exceeded when he or she publicly displays this level of participation. Actant citational types like author, then, may point to additional in-depth text mining work done on a larger corpus, which would help corroborate these predictions.

A baseline of fidelity between advisor and advisee writings may be established by comparing the distances between advisor and advisee writings, we believe, which might shed insight on how advisees are learning about citation practises and how advisors are modelling citation practises. Genre convention mastery is and will be a moving target. Citational motions may be required for publication, but they may also be required in an overly strict manner, which may lead to inconclusive results or hamper the development of new ideas. The network graph measurements offered by our team may be utilised to automate the selection of a suitable general baseline for citational alterations among academic mentoring ties. Using a pairwise metric can serve as a global indicator that an advisee is maintaining the proper citation patterns necessary for field recognition and the preservation of that advisee's scholarly voice after an advisor has established a qualitative judgement about the acquisition of disciplinary writing codes. Aside from the ability to track citations in various academic contexts, such as academic journals, dissertations, and academic books, we also want to be able to track the emergence of new fields through the usage of citations in the future.

Notes

1. The four-part citation coding technique proposed by Moravcsik and Murugesan (1975) contains the following:

a. conceptual (the citation refers to a theory or concept) or operational (the citation refers to a method or procedure) (citation references a method)

b. organic (essential for understanding the current or cited paper's content) or perfunctory citation (acknowledgement of previous work)

c. evolutionary or juxtapositional (current paper builds on previous work) (current works provides an alternative to referenced work)

d. affirmative (the current study confirms the work of the referred work) or negative (the current publication refutes the work of the referenced work) (current paper challenges or critiques referenced work)

2. 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 (referenced work or findings are integral to understanding the referenced work but not related to the content of the referring paper) d. supplementary (referenced work or findings are integral to understanding the referenced work but not related to the content of the referring paper) (references provide additional, independent information)

d. insignificant (references included with interpretation)

e. partial negation (references that the referring article disagrees with in part)

f. total negation (references that the referring article rejects outright).

3. See Cronin 1984, pp. 35-49, for a discussion of citation classification taxonomies.

4. Teufel et al. (2006)'s study expands on previous categorization attempts aimed at automatically diagnosing the content of research articles by extracting information about the rhetorical status of phrases. The assignment of rhetorical status on a sentence-by-sentence basis may enable better automatic text summarization (Teufel and Moens, 2002) and more informative citation indexing (Teufel 2006).

5. The text processing technique used in the Springer OpenAccess research database to categorise citational and non-citational sentences is based on the use of predefined regular expression criteria. http://ryan- omizo.com/2016/01/11/finding-genre-signals-in-academic-writing-benchmarking-method/

6. In Paul's (2000) study, location is likewise used as an operant citational attribute.

7. In the next 30-75 percentile, 30.2 percent of citations emerge; in the final 75-100 percentile, 22.6 percent of citations appear.

8. We use sci-kit learn's pairwise distance metrics algorithm for this study (see Pedregosa et al. (2011) and scikitlearn's module documentation at http://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.pairwise distances.html).

9. Swales (1990, p. 213) proposes that models be presented in "caricature" format, which "simplifies and distorts" genre features in order to acclimate students to the rhetorical demands of academic writing. This attenuation process is seen as a way to limit the variety 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 inbipolar 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: 10.1162/08912010 2762671936
- 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.