ABSTRACT

This study investigates the use of epistemic stance features within a specialized, diachronic corpus of biochemical research pertaining to the motility of bacterial cells in a process referred to as chemotaxis. The corpus constructed for the investigation includes 328 open access research articles citing the seminal 1972 publication, "Chemotaxis in *Escherichia coli* analysed by three-dimensional tracking" in the peer-reviewed journal *Nature* by Drs. Howard Berg and Douglas Brown. For the investigation, the corpus was segmented into sub-corpora representing five time periods and the trends in use of epistemic stance markers were analyzed. Over the period covered by the corpus (1972-2017), the overall use of modal auxiliaries and non-modal hedges decreased while the frequency of boosters increased. Additionally, epistemic stance markers indexing greater degrees of certainty increased while epistemic stance markers reflecting doubt and uncertainty decreased. These findings are noteworthy as they contrast with previous studies investigating diachronic change in epistemic stance use in both academic and science writing and add to our understanding of the use of epistemic stance in the formation of knowledge.

Keywords: *diachronic analysis, epistemic stance, science writing, corpus-aided discourse study*

Epistemic Stance and the Construction of Knowledge in Science Writing:

A Diachronic Corpus Study

Few domains of discourse can yield as much insight into the formation of knowledge as science writing, and thus, scholars from a range of disciplines from applied linguistics to rhetoric and sociology have long endeavored to understand the complex process through which a claim or proposition is first presented and then later accepted as fact. While some may continue to perceive professional science writing as a rather distant and objective reporting of truth and fact, it is widely acknowledged and documented that science discourse is both socially situated and constructed (Bazerman, 1988; Latour & Woolgar 1979; Hyland, 1996). Halliday and Martin (1993) go so far as to call for abandonment of the rather "naive" notion that science writing simply reflects an objective, external reality and instead implore scholars to adopt a social constructivist approach to the language of science (p. 8). As Hyland (2005) would later note, "Rather than simply reporting studies of the natural or human worlds, writing actually helps to create a view of these worlds...in other words, texts cannot be seen as accurate representations of what the world is like because this representation is always filtered through acts of selection, foregrounding, and symbolization; reality is constructed" (p. 141). Nonetheless, the traditional perception of science writing as objective, distant, and impersonal persists, a perception some believe “a very dangerous myth" (Crismore & Farnsworth, 1990, p. 118).

On a macro-level, a view of science as a rational and cumulative endeavor in which knowledge accumulates in a sort of "piecemeal process" by which facts and methods are added to an "ever-growing stockpile" of accepted knowledge is not wholly inaccurate (Kuhn, 1962, p. 1-2). It is at the micro-level of metadiscourse where evidence of the socially constructed

qualities of science writing may be seen. And while science writing has shifted towards fewer overt markers of authorial presence (Atkinson, 1999), interactive and interpersonal features, i.e. metadiscourse, remain quite frequent in science writing, although it should be noted that overt interpersonal features are less frequent in academic discourse than in most registers (Gray and Biber, 2012).

Insights into the social qualities of science writing have emerged from analyses that identify changes in the use of linguistic features indexing authorial presence and the narrative style of early science writing (e.g. Atkinson, 1999); this narrative quality sharply contrasts with the contemporary style of science writing and its preference for abstractness and information density (Atkinson, 1999), the style Halliday and Martin refer to as "the language of science" (1993, p. 2). This discursive evolution has been highlighted in the publications of the Philosophical Transactions of the Royal Society of London from 1675-1975 through corpus-driven multidimensional analysis. Similar diachronic investigations from rhetorical studies have further highlighted the social qualities of science writing and explored how these social features help authors align with conventions and norms of their community while forwarding findings and seeking consensus. Notably, Bazerman (1988) analyzed peer-reviewed journal articles pertaining to spectrometry between the years 1893-1980, asserting that knowledge claims within scientific texts are "highly contextualized linguistic products, the printed trace of complex systematic activities" (p. 153), and notes how authors "control the language and presentations of their papers so as to present their work in the most favorable light, so as to advance the acceptance of their own work, and to further their interests as scientists" (p. 156). Thus, science is "realized only through linguistic, rhetorical, and social choices, all with epistemological consequences" (Bazerman, 1988, p. 183), and "the act of accrediting knowledge is a social

process and involves making linguistic choices" (Hyland, 2005, p. 143). Through an analysis of epistemic stance in a specialized corpus pertaining to one particular domain of biological research, this study seeks to extend understanding of epistemic stance features and their functioning in the construction of knowledge within a specific discourse community and its focus of inquiry.

Epistemic Stance

Stance has received much attention from applied linguistics with various terms and approaches applied to its study. Defined as the "attitudes, feelings, judgments, or commitment concerning the propositional content of a message" (Biber & Finegan, 1989, p. 93), stance is realized through a range of grammatical devices from modal auxiliaries (e.g. *may, might, can*) to adverbial hedges (e.g. *probably, actually, likely*) as well as certain verbs (e.g. *seems, suggests, demonstrates*). As indicated, stance markers function to "convey many different kinds of personal feelings and assessments, including attitudes that a speaker has about information", and importantly for the purpose of this study, "how certain they are about its veracity" (Biber, 1996).

While this study adopts the stance framework and the definition of stance offered by Biber and Finegan (1989), other terms for the constellation of lexicogrammatical features that index and reflect attitudes and opinions towards various claims or propositions have been forwarded, perhaps most notably evaluation (Thompson & Hunston, 2000) and appraisal (Martin, 2000). Whether stance, evaluation, or appraisal is the preferred and selected superordinate term, these terms are linked by their common interest in exploring how various grammatical devices reflect an author's beliefs and values, how these linguistic features help authors manage relationships with their readers, and how these items enable writers to produce effective texts within their discourse community.

One particularly productive area of stance research has been in the area of written and spoken academic registers with much attention given to how various grammatical features both overtly and covertly imbue academic writing with interactional and interpersonal qualities (e.g. Biber, 2006; Gray & Biber, 2012; Hyland, 1996a, 1996b, 1998, 2005a, 2005b, 2010). Prior work has detailed the presence of stance markers across academic registers, e.g. undergraduate writing (Aull & Lancaster, 2014), graduate student theses (Charles, 2006), and a range of spoken and written academic registers from textbooks to lectures (Biber, 2006). Additionally, stance features have been investigated in science writing, e.g. hedging in research articles in molecular biology (Hyland, 1996) as well as stance markers in results sections of experimental research articles by biochemists (Thompson, 1993). Often reporting variation in the function and types of stance deployed, Hyland (2005) notes that linguistic devices that academic writers utilize are "community-sensitive" (p. 190). Thus, research into stance use facilitates and enables "the uncovering of the rhetorical and social distinctiveness of disciplinary communities" (2004, p. 156). Stance features in science writing reflect alignment with the conventions and expectations of a community by tempering the strength of claims in respect and even deference to a community of esteemed colleagues (Hyland, 1996, p. 434). Importantly, the presence and function of stance devices highlight the rather interactive quality of science writing and the conventions that must be heeded for knowledge claims to be ratified and adopted by a community (Hyland, 1996).

Within the study of stance, there is a dichotomy produced between epistemic and attitudinal stance with investigations often focusing on one of these domains. This distinction is determined by whether a device reflects personal attitudes and feelings (e.g., *fortunately, interestingly, or luckily*) or if the marker indicates a speaker/author's relationship to the

information detailed in a claim or proposition (Biber, Johansson, Leech, Conrad, & Finegan, 1999); epistemic and attitudinal stance are also referred to as evidentiality (Chafe, 1986) and affect (Ochs & Schieffelin 1989). Epistemic stance devices reflect and index meanings of certainty, doubt, actuality, precision, or limitation (Hyland & Jiang, 2016) and are frequently realized in academic discourse through a range of devices, e.g. evaluative lexis, grammatical choices, and paralinguistic features (Biber, 2006). While evaluative lexis is increasingly studied through corpus-aided approaches, stance has most often been investigated through the analysis of overt grammatical choices such as modal auxiliaries such as *may* and *might*, adverbials like *actually* and *certainly*, or verbs such as *feel* or *believe*. These devices are categorized as hedges when indicating a more tentative position and boosters when signaling confidence in the accuracy of a claim (Hyland & Jiang, 2016). While stance features have a range of socio-pragmatic functions, one affordance these stance devices provide an author is the ability to present claims to an esteemed audience/readership with "appropriate accuracy, caution, and humility as he/she seeks approval and acceptance of his/her claims" (Hyland, 1996, p. 134).

Though much attention has been extended to stance, its realizations, and its various socio-pragmatic functions, much less has been offered to diachronic changes in its application in particular disciplines or in relation to particular claims. As noted previously, diachronic changes in scientific writing have displayed the emergent and social qualities of science discourse, namely Atkinson's study of articles published in Philosophical Transactions between 1675 and 1975 that noted a movement towards a more abstract, informational, "object-centered" style. More recently, diachronic changes in the deployment of stance features such as hedges, boosters, and self-mention have been investigated (Hyland & Jiang 2016); the authors analyze the diachronic changes in frequency of stance markers over the previous five decades across four

disciplines. Hyland and Jiang (2016) shed light on changes in the realization and frequency of stance in research writing and assert the presence of "slow changes in traditional knowledge construction practices" with a decrease in overt stance expression in "more discursive fields" such as sociology and applied linguistics in contrast to a rise in stance features in biology and electrical engineering (p. 269). As Hyland and Jiang note, this change in authorial stance within science writing reflects similar claims forwarded by Atkinson (1999).

The aforementioned analyses of stance markers across academic registers have generally been synchronic investigations meant to report and describe how stance markers perform various functions within an academic discourse community at a given moment in time. These studies have reported how scholarly writers modulate the strength of their claims, but to our knowledge, the study of diachronic change in epistemic stance expression in science writing is largely unexplored, though Hyland and Jiang (2016) is a noteworthy exception. The present study, however, contrasts rather sharply from this previous work on diachronic change in stance in that instead of quantifying the overall frequency of stance expressions in a particular register or in texts from a particular discourse community, we analyze stance features and their impact on knowledge construction in a specialized, diachronic corpus of texts pertaining to one specific topic, chemotaxis, in an attempt to explore how realizations of stance in a specific domain evolve through time and whether these patterns reflect previously reported changes in stance features from more general studies. Our findings diverge from previous studies of stance and give insight into how these salient features function in science writing.

Background: The Study of Chemotaxis

Biochemical and biophysical theory and experimentation seek to understand relationships in biological systems (e.g. cells) through the application of chemical and physical theories. Fields

drawing from such theory and experimentation are therefore interdisciplinary, and, moreover, experimentation and theory in biological chemistry are often informed by biological physics and vice versa. The interconnectedness of biological, chemical, and physical study is evidenced by the number of disciplines that exist in subtle distinction from one another (e.g., molecular biology, biochemistry, biophysics, physical chemistry). One domain of biological research that has been greatly informed by biochemical and biophysical investigations is microbiology, the study of the lives and behaviors of the smallest organisms.

Researchers in this domain have discovered that certain types of bacteria swim in an attempt to change their chemical environment, either escaping their own waste chemicals or seeking nutrients. This behavior is called chemotaxis and falls into the larger class of phenomena termed cellular motility. Chemotaxis then describes the motions that organisms undergo in response to chemical stimuli. Though bacterial motility, including chemotaxis, was explored prior to 1972, the development of an innovative microscopy technology by Drs. Howard Berg and Douglas Brown allowed more robust observations of bacterial motility, which greatly contributed to scientists' understanding of the rather complex set of phenomena it comprises. In other words, Berg and Brown actually invented the microscope that permitted their breakthrough observations. As is often the case in scholarly and scientific research, a breakthrough method and new findings promoted a proliferation of research that elaborated upon, further assessed, and extended the findings of the initial publications.

The Berg and Brown (1972) article has had immense influence on the study of chemotaxis, having been cited approximately 2,000 times according to Google Scholar and ranking as one of Berg's most influential publication on Semantic Scholar. Their methodological innovation enabled the tracking of bacterial cells in three dimensions, a development that

permitted the observation of bacterial behavior in response to external stimuli (e.g., gradients of chemical concentration) at an unprecedented level of detail. The 1972 paper introduced the interplay between two bacterial motions, which the field ultimately agreed to call runs and tumbles. Drawing on physical theory, Berg and contemporaries began characterizing the types of motions that bacteria perform and how these motions give rise to coordinated motion of the cell and are controlled by the cell's chemical environment. Research directed towards cultivating a more comprehensive understanding of bacterial motility has spanned many subdisciplines of biology, chemistry, and physics, and indeed hundreds of articles, many citing the seminal 1972 publication, have been published on the topic.

Corpus and Methods

The corpus compiled for this study includes exclusively open access experimental research articles citing the 1972 publication in volume 239 of *Nature* titled "Chemotaxis in *Escherichia coli* analysed by three-dimensional tracking". We identified research citing our seed article by querying both Google Scholar and Semantic Scholar[1] and retrieved bibliographic information for each publication. Constraining our analysis to open access content, we normalized the title of each such publication retrieved from Google Scholar, and queried against titles in Semantic Scholar's Open Academic Corpus (Ammar et al., 2018). This imperfect method of record matching was necessary, as no citations from Google Scholar include a Digital Object Identifier (DOI). We supplemented results culled from Google Scholar with those referencing publications indexed by Semantic Scholar using their publicly available API. For entries in this set where a URL to a PDF of the work was available, we extracted the paper's full text using version 1.2.8 of Science Parse[2]. When only the abstract was indexed by Semantic

[1] http://api.semanticscholar.org/
[2] https://github.com/allenai/science-parse/releases/tag/v1.2.8

Scholar, we manually searched for the article and collected the text for inclusion into the corpus for analysis.

It is important to emphasize that the corpus, henceforth referred to as the Chemotaxis Corpus, is populated only with open access publications; thus, while approximately 2,000 articles cite our seed article, the Chemotaxis Corpus includes 328 open access publications from 1972 to 2017. As open access publication has only become commonplace in science writing in recent years, the sub-corpora representing the early years of writing on the topic are much smaller than sub-corpora from the year 2000 to the present. Details of the Chemotaxis Corpus are included in Table 1.

TABLE 1: CHEMOTAXIS CORPUS

The epistemic stance features under analysis were modal auxiliaries, hedges, and boosters. Central modal auxiliaries such as *may, might,* and *can* are often included within the taxonomy of hedges, but due to their high frequency and functional saliency, they are scrutinized separately from the other items in the present study. Hyland (2005) and Biber (2006) were consulted in order to generate word lists of epistemic features for automated extraction from the corpora. Ultimately, 8 central modal auxiliaries, 77 hedging devices, and 54 boosters were analyzed.

As noted previously, the size of the corpora from early decades is limited due to the inability to source open access articles from these periods. Thus, as is common practice in corpus-aided discourse studies, the frequencies of occurrence of the items were normalized to per 100,000 words to enable comparison of the five time periods. Additionally, following Hilpert and Gries (2009), Kendall's τ correlation coefficient was calculated to facilitate further statistical evaluation of the strength of diachronic trends. In addition to the initial quantitative analysis, a

post hoc qualitative analysis of various features and patterns deemed salient was also conducted to further explore and contextualize quantitative findings.

Results

Modal Auxiliaries

Modal auxiliaries are a key linguistic feature reflecting and indexing epistemic stance, and thus, an analysis of the eight central modals (*can, will, may, must, could, might, would, should*) is the entry point into the inquiry. It should, however, be noted that depending on the context, a modal may perform either a hedging or boosting function. Broadly, in the full Chemotaxis Corpus, modal auxiliaries account for 20,935 tokens with the possibility/ability modal *can* (9340 instances) far exceeding the second most frequent item, the prediction/volition modal *will* (2875 instances). A closer analysis of modal frequencies over the five time spans revealed several insights: 1) the normalized frequency of use of 6 of the 8 central modal auxiliaries exhibited an overall decrease over the time periods represented by the sub-corpora, 2) *can, will,* and *may* are the most frequently occurring modals of the five time periods 3) the two modals indexing highest degrees of certainty, *can* and *will* are the two modals whose use is increasing.

TABLE 2: MODALITY ACROSS TIME PERIODS

An analysis of the normalized frequencies of the modal auxiliaries over the five time periods reveals clear use trends for the items. For 6 of the 8 central modals, the normalized frequencies are decreasing with frequency numbers increasing for only *will* and *can.* The modal *can* is classified as a modal performing functions of possibility, permission, and ability; permission is not a common function for *can* in the Chemotaxis Corpus. Within the corpus, it typically expresses either possibility or ability reflective of the instances in samples a, b, c, and d

extracted from the corpus, though it is sometimes rather difficult to definitively distinguish possibility and ability.

a) Here we show that we ***can*** accurately account for the complex and unique crawling patterns exhibited by individual *Drosophila* larvae using a small set of four parameters obtained from the trajectories of a few crawling larvae.
b) The transmission spectrum shows that even at peak absorption wavelengths, a single cell ***can*** absorb only about 20% of the photons that pass through it.
c) Some motile cells ***can*** change their moving directions in response to chemical environment changes.
d) At low fluid velocities, it is possible that bacteria ***can*** accumulate at the capillary mouths by migrating upstream in the flowing bacterial suspension in response to the chemoattractant gradient that forms downstream of the capillaries.

These four instances of *can* are reflective of its primary functions within the corpus; the four instances are provided on a general cline of certainty to uncertainty and, thus, less to greater potential rhetorical risk to the author. In the first instance, the epistemic verb *show* is followed by a *that*-clause whose use of the pronoun *we* reflects the author's conviction in the claim and willingness to be directly connected to the statement. Additionally, the ability modal *can* is followed by the adverbial *accurately* that further imbues the statement with strength. In contrast, the second statement includes no self-mention or additional epistemic indicator beyond the use of the ability modal *can.* The third statement (c) is somewhat similar to (b) yet the adjectival *some* constrains and places a boundary on the claim. Finally, the final sentence (d) displays the greatest degree of reservation and tentativeness as the proposition containing the modal is hedged with the epistemic adjective of likelihood *possible.* Though these few sentences reflect

general use patterns of the modal *can,* the use of this ability/possibility modal was also scrutinized when used in the collocational frame with the term of the phenomenon at the core of the study: chemotaxis.

Considering the high frequency of *can* amongst the core modals, it is unsurprising that the *chemotaxis* + *can* bi-gram represents approximately 40% of the total *chemotaxis* + *modal* pairings. Additionally, all *chemotaxis* + *modal* bi-grams include modal auxiliaries indicating functions of ability/possibility. For each of the five time periods, *can* is both the most frequent modal and the most common modal auxiliary following the term *chemotaxis* in the corpus. It should be noted, however, that the bi-gram *chemotaxis* + *is* greatly exceeds the *chemotaxis* + *modal* pattern. The *chemotaxis* + *is* patterns unequivocally define the parameters of the phenomenon of chemotaxis and reflect the certainty of the field in describing and explaining the long-studied phenomenon. With such variety of choice within the language system, what is the functional distinction and rhetorical purpose of the contrasting pairs *chemotaxis* + *is* and *chemotaxis* + *modal*? The selection denotes a boundary of knowledge and the researcher's first step into uncharted space. In other words, the *chemotaxis* + *is* patterns establish consensus and reflect existing knowledge about chemotaxis while the *chemotaxis* + *modal* pattern seems to expand the field and open a research space for continued study. Many of the *chemotaxis* + *modal* patterns manifest the potential for future research as the hedged claim provides the opportunity for additional research questions and continued investigation. Thus, the approximately 40 *chemotaxis* + *modal* patterns either emphasize community knowledge or delineate a boundary marker on a knowledge map signaling liminal spaces to be pushed and challenged.

The following samples display patterns with the *can* ability modal indicating shared knowledge space while the subsequent *may* samples reflect the possibility function and open new

research trajectories. In the first sample, *can* operates in a statement of fact and reflects a shared understanding of a research-attested behavior. Similarly, the second instance makes a factual statement regarding the process of measurement while the final *can* sample affirms a basic function of chemotactic behavior established within the community. In contrast, the uses of *may* hedge the forwarded knowledge claims, thereby recognizing an untested possibility and opening a research space. Though *may* and *can* could potentially be equally suitable choices in some settings, their uses in this community seem to diverge due to the certainty encoded within *can* that is absent in the modal *may.* While *may* is more ambiguous in its functioning, *can* is more definitive and the certainty it indexes allows the author to align with established knowledge structures.

- The behavior underlying chemotaxis ***can*** be evoked by adding attractant to bacteria and observing that they swim smoothly…(Ordal, 1976)
- Chemotaxis ***can*** be measured by the formation of rings of bacteria around the attractant plugs. (Poole, 1989)
- Chemotaxis ***can*** enhance the rates at which microbes exploit nutrient patches. (Lee et al., 1999)
- Therefore, bacterial chemotaxis ***may*** provide a basis for the search of a global optimum. (Muller, 2002)
- Run-and-stop chemotaxis ***may*** be ineffective and disadvantageous where a gradient is moving. (Mitchell, 2002)
- Bacterial chemotaxis ***may*** have a significant impact on the structure and function of bacterial communities. (Law, 2005)

Following the directive of Hilpert and Gries (2009) for the analysis of diachronic corpora, Kendall's τ was calculated to enable a closer statistical inspection of use trends beyond simple eyeballing of normalization data. This measure was calculated using the relative frequencies of the 8 modal auxiliaries over the five periods represented in the corpus. For interpretation of the data, a value near 0 reflects that no trend is present (i.e. irregular rises and falls across time periods) while values approaching -1 or 1 indicate that the differences in relative frequency of a lexical feature observed across diachronic sub-corpora are negative (decreasing in time) or positive (increasing in time). In other words, when τ approaches 1 or -1, "the passage of time correlates perfectly with an increase or decrease in frequency" (Hilpert & Gries, 2009, p. 390). If an item's frequency fluctuates yet overall there is an increase in use, the value will be positive but not approaching 1. Similarly, if an item generally decreases yet experiences "peaks" and "troughs", the value will be negative but not approaching -1 (Gabrielatos, et al., 2012, p. 1). Though scrutiny of the normalized frequency data does provide insight into use trends, the statistical evaluation made possible through the Kendall's τ enables analysis of the strength of trends for the modals. The Kendall's τ values indicate positive use trends in use for three modals (*can, will* and *need*) and negative use trends for the remaining items. Additionally, the values for *will* and *need,* though positive, do not reach a level of statistical importance. Of the items returning a positive Kendall's τ value, only the increase for *can* occurs at a rate of marginal significance at 0.600.

While *can* occurred at the highest frequency but with a weak Kendall's τ, the modals *would, should,* and *might* recorded scores closely approaching -1.0, with *could* closely behind. In total, four modal auxiliaries have negative use trends in closer proximity to -1.0 than any of the positive modals appear to +1.0. Hyland's (1998) analysis of hedging in scientific writing notes

that this content-oriented hedge functions in "mitigating the reliability of the conclusions by making them dependent on the validity of the premise" (p. 198) and suggests that *would* adds to the "persuasive character of the argument by deferring to the reader in emphasizing potential correspondence of results to an actual situation in the external world" (p. 198). However, in the Chemotaxis Corpus, the use of *would* is decreasing at a significant rate, leading one to conclude that the need to hedge and cautiously forward claims is less urgent and necessary as the shared body of knowledge regarding chemotaxis has perhaps solidified. With less need to forward claims in a manner that builds consensus and asserts propositions with caution, the rhetorical function of *would* is less prominent in this evolving domain, and thus, its frequency of use is steadily declining.

The modal *must* was selected for deeper analysis due to the multiple discursive functions it may perform and its generally decreasing trajectory of use. Though Biber (2006) assigns meanings of necessity and obligation to *must,* a concordance analysis of a sample of its occurrences within the Chemotaxis Corpus motivated the assignment of three functional codes for analysis. The three functions are: 1) deduction/logic: instances where a logical deduction is being offered by an author, 2) obligation: requirements for adherence to certain processes and procedures often when discussing methodology, and 3) knowledge: general statements of knowledge often referencing community knowledge. Though the codes were discussed as a unit, the team member with experience in biochemical research coded a random sample of 100 instances of *must*; each time period was represented proportionally in the sample. It was concluded the insider status and knowledge possessed by this team member would allow more accurate coding of the 100 instances in accordance with the three nuanced functions represented by the codes. An exemplar of each code is provided below.

- Deduction/Logic: These results indicate that decisions ***must*** overlap in genetically silent regions between the tar and cheR genes (Slocum, 1983)
- Obligation: A maximum estimate ***must*** consider that the cell may not move from the initial orientation to the final orientation in a straight line. (Mitchell, 2002)
- Knowledge: To survive, a *Bdellovibrio* ***must*** locate and successfully penetrate a prey cell before losing its viability due to rapid starvation. (Straley, 1997)

Though the distinction between knowledge and obligation is rather subtle, the instances of *must* marked as knowledge indicate instances of shared and accepted community knowledge while the obligation instances report methodological requirements and/or procedural decisions. Such a distinction would be difficult to discern by a non-expert reader, thus reflecting the context-sensitive, social nature of stance in a particular discourse community. The deployment of *must* in these varied functions indexes membership in the discourse community through alignment with existing knowledge of the field as well as adoption of procedural measures. However, it is likely that through time such overt expressions of community-accepted knowledge would be less likely while use of the modal in methodological contexts would remain necessary. These two functions and their divergent nature of use likely explain the irregularity in its use through the time periods. Though a deeper analysis of each function throughout the five time periods would be beneficial, such work is beyond the scope of the present study.

It seems plausible and largely intuitive that the frequency with which modals are deployed in relation to a particular proposition would decrease as time passes and understanding of a phenomenon gains clarity. If doubt persisted regarding the process of chemotaxis, the frequency of modal auxiliaries would likely increase as researchers and scientists continued to consistently hedge claims and tentatively offer explanations for observations in their laboratories.

The data indicate that the modal indexing the highest level of certainty is the one on the rise, while the modals reflecting greater doubt are decreasing. Though the present study and Hyland and Jiang (2016) have several meaningful corpus design distinctions, the present work tends to diverge from the findings of their diachronic analysis. While they note a general increase in the use of modals in general writing of the hard sciences, the present study adds nuance to this finding as the modal use in the Chemotaxis Corpus seems to evolve as the community's understanding and agreement regarding a particular phenomenon reaches greater consensus.

Hedges

In addition to modal auxiliaries, hedges were also extracted and analyzed across the five time periods of the Chemotaxis Corpus; modals were retained in this step only to enable a cumulative count and analysis but are not included elsewhere in the analysis. Notably, the normalized frequency of total hedging devices displayed a decrease in use over the five time periods; this is even with the inclusion of the high frequency modal *can* trending towards greater use. Additionally, as was done with modal auxiliaries, the Kendall's τ value was calculated to statistically evaluate the trends of hedging devices. For overall hedge use in the Chemotaxis Corpus, the Kendall's τ value was -0.800. This indicates a rather clear, consistent, and statistically meaningful downward trend in the use of hedging devices in the Chemotaxis Corpus from 1972 to 2017.

Due to considerations of space, a Kendall's τ value of +0.6 and -0.6 was established as a cutoff to narrow the analysis to items displaying particularly strong and clear increasing or decreasing trends in use over time. Though many hedges exhibited Kendall's τ values approaching 0, indicating that their use remained essentially consistent or fluctuated in use over time, 28 hedge words (not including modals) exhibited clear increasing or decreasing trends

beyond the +0.6 and -0.6 cutoff points. Ranked from +1.0 to -1.0, Table 3 exhibits that only 8 items experienced clear increases, as reflected in positive Kendall's τ approaching +1.0, while greater than twice that amount exhibited steady decreases in use. To clarify, of the 77 non-modal hedging devices extracted and analyzed, only 8 items exhibit increases in use, 49 reported frequencies that were either consistent or irregular, and the remaining 20 had clear decreases in use.

TABLE 3: HEDGES OVER TIME

When discussing modals, it was noted that the modal of highest frequency *can* also indexes greater certainty and was the item trending most sharply towards greater use. Though hedges do signal a tentativeness towards a claim, it is also reasonable to assert that the degree of tentativeness occurs on a cline. Qualitatively, it appears a fair observation to note that many of the hedges whose uses are trending positively towards a +1.0, e.g. *often, typical, frequently, probably*, are epistemic adverbials indexing higher levels of certainty towards a claim. These may not appear in purely declarative contexts that express absolute certainty, but they are moving more towards closure and certainty regarding the mechanisms of chemotaxis. As was displayed with modals in the previous section, it seems that as the study of chemotaxis matures and develops the authors' exigency to hedge declines. As space is constrained, only hedging adverbials *often* and *probably* will be explored in greater depth.

The epistemic adverbial *often* does not indicate closure but does signal a trend towards a degree of confidence regarding a claim and proposition. It also indicates alignment with accepted knowledge in the field. A collocational and concordance analysis of *often* revealed a common functional pattern produced through *it is often referred to/used to/leads to/attributed to.* These patterns index the author's alignment with existing conceptualizations and understandings of the

field. In contrast, the epistemic stance adverbial *probably* appears in patterns in which the author assumes a more tentative, speculative tone. The common patterns for this adverbial are *probably due to, probably because, probably result from.* As present with modals, the item that tends to index a greater degree of certainty and confidence is trending to greater use while items trending negatively are those that reflect uncertainty and doubt. This finding suggests that as a field's understanding of a phenomenon matures, the types and frequencies of hedges also evolve.

The decrease in the frequency of hedges reflects the use patterns present with modal auxiliaries. As researchers gain confidence in their understanding of chemotaxis and bacterial motility, one would expect the need to modify claims with certain modals, temper assertions with various adverbial hedges, or modulate statements with adjectives such as *plausible* or *apparent* to decrease. No longer, or at least with much less frequency, would authors need to produce patterns such as *plausible ranges, plausible mechanisms*, or *a plausible model* as well as *probably a result of*, *probably due to*, or *probably because of*.

Boosters

A final category comprising lexical boosters was also extracted and analyzed. With the overall use of hedges on the decline, one would likely expect a contrasting trend from boosters. This expectation was confirmed, and indeed the strength of the trend for boosters exceeded that of hedges. As noted previously, the Kendall's τ correlation coefficient was -0.800 for cumulative epistemic hedge use; in contrast, the value was +0.949 for boosters. This value indicates an increasing trajectory of boosters in the literature pertaining to chemotaxis and bacterial motility. A closer study revealed that, in the first two sub-corpora, there were no occurrences of many of the boosters. Overall, the normalized frequency for total booster use grew from approximately 30 tokens per 10,000 words to 40.

As was present with hedges, there were some irregular variations in the frequencies of many boosts, i.e. individual lexical items may have started at one point, experienced a fall/rise, then exhibited more consistent use patterns. Nonetheless, there were items displaying defined and consistent use patterns. It should be noted as well that while greater than 80 lexical boosters were included in the analysis, the list of boosters observed includes roughly half that figure. With similar cutoff points applied as with hedges, only five boosters reflected Kendall's τ values above/below the +0.6 and -0.6 thresholds. This indicates two possibilities: 1) many boosters are occurring at similar frequency rates across the time periods, and 2) there is much greater fluctuation, i.e. more rises and falls in the use of boosters than with modals and hedges. The normalized frequencies of several high frequency boosters display the greatest variation amongst the boosters. For example, for the 2012-2017 period, the booster *show* enjoyed the highest normalized frequency. However, *show* was used at a rather similar rate between 1972-1981, fell sharply in the 80s, and later rebounded to its present high level of use. It should be noted, however, that although *show* is often considered a booster, it lacks the strength of similar items such as *demonstrate* or *prove.* An additional high frequency booster *found* had similar irregularity over the time periods.

TABLE 4: BOOSTERS OVER TIME

Discussion

Hyland and Jiang's (2016) analysis of stance features in peer-reviewed journals representing four academic disciplines (applied linguistics, biology, electrical engineering, and sociology) from the years of 1965, 1985, and 2015 is significant for its diachronic analysis of stance features that are most often treated from a synchronic perspective. Their study discovered a consistent increase in stance features in academic writing in general with the total use of stance

items increasing by approximately 50% from 1965 to 2015. More specifically, they found steady increases in the use of hedges in engineering and biology along with a consistent decrease in boosters within biology, a field closely connected to the domain explored in this study. Importantly, the modal hedge *may*, an auxiliary indexing an ambiguous and rather tentative degree of certainty, was the most frequent hedge in the two sciences, though its use was slightly decreasing. In sharp contrast, the present study revealed a decrease in hedges and discovered *can* to be the most frequently occurring modal hedge. Our study indicates divergent epistemic stance use between a general corpus of science writing and a specialized corpus of science writing focused on one topic. This divergence between the use of stance markers in academic writing generally compared to presence of stance features in a focused area of research represents an important finding that adds nuance to our knowledge of stance.

The decreasing use of epistemic hedges revealed in our analysis corresponds to the development and maturation of a particular domain of study. While a field would likely experience increases in the use of modals and other hedging devices for a certain period of time as new questions are asked, novel research pursued, and fresh observations shared, the data in the present study reflect an eventual closure, although a new line of inquiry could again necessitate renewed hedging or modality in this domain. As understanding is achieved and consensus attained, the frequency with which hedging devices occur in the explanation of a particular phenomenon decreases over time. More clearly, the frequency of epistemic devices indexing uncertainty and doubt decreases as a particular domain of study develops and matures; simultaneously, the frequency of epistemic devices indexing greater certainty for a particular phenomenon increases. Thus, while epistemic stance markers are a highly functional element of science writing whose use will certainly persist, our data indicate that the types and frequencies

of hedges deployed evolves as a community reaches consensus on its understanding of a phenomenon, an insight made possible through our topic-focused, specialized corpus.

Our study displays how epistemic modal auxiliaries and hedging devices intended to mitigate potential rhetorical risk while balancing social and disciplinary conventions and norms decline as the knowledge base pertaining to a particular phenomenon solidifies. Though difficult to identify, there is a point where a community accepts a proposition as a fact and thus reduces its tendency to hedge claims regarding its features and mechanics. Thus, it seems reasonable to conclude that author presence as reflected in epistemic stance features becomes less overt as a discipline adopts a shared understanding of a phenomenon. As Hyland (2016) notes, "stance is always expressed in relation to some set of expectations so that personal judgements are convincing" (p. 254). In regards to chemotaxis and motility, the need for and expectation of stance markers to index whether explanations are valid, interesting, and useful to the community are less urgent as the field advances.

This study exhibits the additional value of exploring epistemic stance more narrowly through the investigation of specific phenomena beyond more general investigations. It seems reasonable to anticipate the sort of evolution in epistemic stance observed in the present study. Disciplines are ever expanding, evolving, and entering new domains of inquiry while asking ever more complex and sophisticated research questions. As researchers and writers chart their path into yet unexplored research frontiers, the need to tentatively assert claims certainly is not yielding. It is logical and expected that authors presenting novel claims in uncharted areas would continue to deploy high frequencies of epistemic hedges, and, as our data affirm, it is equally anticipated that as the knowledge solidifies those same epistemic hedges would decrease as boosters increase. Our quantitative findings affirm such expectations while opening additional

paths for research. More research into specific knowledge structures surrounding specific phenomena would contribute to a greater understanding of epistemic stance and its functioning in the construction of knowledge.

Limitations

Though the Chemotaxis Corpus did yield insights into the diachronic formation of knowledge, we acknowledge several limitations of the present study. Most notably, the dearth of open access articles available for the first two time periods in the Chemotaxis Corpus is an issue. Though minimum thresholds for corpus size are somewhat arbitrary, most corpus linguists would agree that the size of the first sub-corpus is rather small. To mitigate this issue, data was normalized to allow comparison of the sub-corpora and relative frequencies were calculated for the calculation of the Kendall's τ values; these are both customary procedures in corpus-aided discourse studies that explore variation in differently sized corpora as well as corpus studies investigating change over time. Though we acknowledge that increased size of the sub-corpora of the first two time periods would be beneficial for the study, we believe in the value of open access publication and hope it will be possible to eventually make the Chemotaxis Corpus publicly available. It is our hope that others will interrogate the corpus we have compiled and produce additional insights.

An additional limitation is that our corpus-based analysis procedure imposed a limited range of linguistic devices for extraction and analysis. It is possible that the means through which epistemic stance is realized have grown increasingly nuanced, covert, and discipline specific. It is of value to note that one member of the research team is a biochemical researcher fully capable of engaging with the scientific content of the research literature comprising the Chemotaxis Corpus. Though we believe his experience and knowledge have aided our data

analysis and discussion, we concede that a corpus-driven approach that allows epistemic devices to percolate and appear organically could yield additional insights into the topic. In future work, analysis could pursue a more corpus-driven approach for the analysis of stance as suggested by Plappert (2017).

Finally, our study primarily focused on the clearest and most statistically significant usage trends of various stance features in the corpus. While our study analyzed clearly visible trends as reflected in Kendall's τ scores, there are likely insights in the "peaks" and "troughs" of the data (Gabrielatos, et al., 2012). Anecdotally, we noted that much fluctuation occurred in the corpus representing 1982-1991 as many of the items displayed marked changes in this period before returning to a more stable use trend. It is possible that, in the rises and falls, additional understanding of knowledge formation and the social and linguistic process through which claims are advanced and adopted could be discovered. A deeper analysis of these irregular usage trends will be the focus of future research but is beyond the scope of the present study.

Conclusion

Stance markers and their importance to academic and science writing are well documented. The present study, however, offers insights that diverge from previous research and add nuance to the understanding of the use of these features in science writing. Most notably, our study reveals a consistent decrease in the use of epistemic stance items indexing uncertainty, likelihood and doubt along with clear increases in boosters marking higher levels of confidence and certainty in a specialized corpus pertaining to the exploration of one particular phenomenon. This trend is most clearly displayed in the high frequency and consistent growth in the modal *can* over the five time periods. It is also worth noting that our study benefited from the insights

provided by the presence of a biochemical researcher on the team. The insights he offered would not have been gained without access to his expertise and insider perspective.

Our study and its findings open several avenues for future research. First, while our study explored one particular phenomenon, it would be valuable to replicate and extend our approach to the study of other phenomena from other discourse communities to triangulate and verify the findings presented here. Additionally, future work should explore in greater detail how the functions of particular stance features evolve through time. Finally, our study primarily examined items displaying clear positive or negative use trends. We believe future work could further explore the irregular rises and falls of various stance items.

Acknowledgements: We would like to offer our sincere gratitude to two anonymous reviewers for the many forwarding comments provided on an earlier version of this manuscript. While their comments were uniformly beneficial, we wish to highlight that the added discussion of the modal *must* was largely the result of comments from one of the reviewers. We are indebted to them both.

References

Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D.,…Etzioni, O. (2018). Construction of the Literature Graph in Semantic Scholar. NAACL-HTL.

Aull, L. & Lancaster, Z. (2014). Linguistic markers of stance in early and advanced academic writing: A corpus-based comparison. *Written Communication*, 31(2), 151-183.

Atkinson, D. (1999). *Science writing in sociohistorical context*. Mahwah, NJ: Lawrence Erlbaum.

Bazerman, C. (1988). *Shaping written knowledge*. Madison, WI: University of Wisconsin Press.

Berg, H. & Brown, D. (1972). Chemotaxis in *Escherichia coli* analysed by three-dimensional tracking. *Nature,* 239, p. 500-504.

Biber, D. & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Interdisciplinary Journal for the Study of Discourse*, *9*(1), 93-124.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (1999). Longman grammar of spoken and written English. Essex UK: Longman.

Chafe (1986). Evidentiality in English conversation and academic writing. In W. Chafe, & J. Nichols (Eds.), *Evidentiality: The linguistic coding of epistemology* (261-272). Norwood, NJ: Ablex.

Charles, M. (2006). The construction of stance in reporting clauses: A cross-disciplinary study of theses**.** *Applied Linguistics*, *27*(3), 492–518.

Crismore, A., & Farnsworth, R. (1990). Metadiscourse in popular and professional science discourse. *The Writing Scholar: Studies in Academic Discourse,* 118-136.

de Marneffe, M. C., & Manning, C. D. (2008). The Stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation* (pp. 1-8). Association for Computational Linguistics.

Gabrielatos, C., McEnery, T., Diggle, P., Baker, P. (2012). The peaks and troughs of corpus-based contextual analysis. *International Journal of Corpus Linguistics*, 17(2), 151-175.

Gray, B. & Biber, D. (2012). Current conceptions of stance. In K. Hyland & C. Sancho-Guida (Eds.), *Stance and voice in written academic genres* (15-33). London: Palgrave-Macmillan.

Halliday, M.A.K., & Martin, J.R. (1993). *Writing science: Literacy and discursive power*. New York, NY: Routledge.

Hilpert, M., & Gries, S. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4), 385-401.

Hyland, K. (1996a). Writing without conviction: Hedging in science research articles. *Applied Linguistics*, 17 (4), 433-454.

Hyland, K. (1996b). Talking to the academy. Forms of hedging in science research articles. *Written Communication,* 13(2), 251-281.

Hyland, K. (1998). *Hedging in scientific research articles*. Amsterdam: John Benjamins.

Hyland, K. (2005a). *Metadiscourse*. London: Continuum.

Hyland, K. (2005b). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies*, *7*(2), 173-192.

Hyland, K. (2010). Constructing proximity: Relating to readers in popular and professional science. *Journal of English for Academic Purposes,* 9, 116-127.

Hyland, K. & Jiang, F. (2016). Change of attitude?: A diachronic study of stance. *Written Communication,* 33(3), 251-274.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago, IL. University of Chicago Press.

Latour, B. & Woolgar. (1979). *Laboratory life: The construction of scientific facts*. Princeton, NJ: Princeton University Press.

Martin, J.R. (2000). Beyond exchange: Appraisal systems in English. In. S. Hunston & G. Thompson (Eds.) *Evaluation in text: Authorial stance and the construction of discourse* (142-175). Oxford: Oxford University Press.

Ochs, E. & Schieffelin, B. (1989). Language has a heart. *Written Communication, 9*(1), 7-25.

Plappert, G. (2017). Candidate knowledge? Exploring epistemic claims in scientific writing: A corpus-driven approach. *Corpora,* 12(3), 425-457.

Thompson, D. (1993). Arguing for experimental “facts” in science: A study of research article results sections in biochemistry, *Written Communication,* 10(1), 106-128.

Thompson, G. & Hunston, S. (2000). Evaluation: An introduction. In. S. Hunston & G. Thompson (Eds.) *Evaluation in text: Authorial stance and the construction of discourse* (1-26). Oxford: Oxford University Press.