Variation in Citational Practice in a Corpus of Student Biology Papers: From Parenthetical Plonking to Intertextual Storytelling

John M. Swales

Written Communication 2014 31: 118
DOI: 10.1177/0741088313515166

Written Communication
2014, Vol. 31(1) 118–141

DOI: 10.1177/0741088313515166
wcx.sagepub.com

Variation in Citational Practice in a Corpus of Student Biology Papers: From Parenthetical Plonking to Intertextual Storytelling

John M. Swales[1]

Abstract

This is a corpus-based study of a key aspect of academic writing in one discipline (biology) by final-year undergraduates and first-, second-, and third-year graduate students. The papers come from the Michigan Corpus of Upper-level Student Papers, a freely available electronic database. The principal aim of the study is to examine the extent of variation in citation practice in the biology subcorpus. To that end, it explores citation practices from a number of perspectives, including the distribution of integral versus parenthetical citations, the choice of reporting verbs, the effect of citing system, and the occurrence of selected features such as the use of citees' first names. Results show little difference between the undergraduate and graduate papers, some effect of the citing system, and a somewhat richer intertextuality in the "evolutionary" as opposed to the "molecular" biology papers. Overall, this is an impressive body of student work from the viewpoint of textual variation in citation practice, but it should be remembered that the corpus consists of only "A" papers from a flagship research university.

Keywords

citations, biology, student writing, corpus, analysis

[1]University of Michigan, Ann Arbor, MI, USA.

Corresponding Author:
John M. Swales, University of Michigan, 546 Fifth St., Ann Arbor, MI 48103-4839, USA.
Email: jmswales@umich.edu

The literature on citations in academic texts is large, much of it coming from information science, but with also sizable contributions from the sociology of knowledge, new rhetoric studies, and English for academic purposes. Work in the first field has tended to be independent of the others, although there have been occasional attempts to bridge the gap (Harwood, 2009; White, 2004). The reasons for this large and complex literature are themselves complex. To start with, citation is the most overt and most immediately obvious indication that a text is indeed academic. Citing permits an author to introduce and discuss the contributions of other researchers and scholars, and through such knowledge displays of previous literatures he or she can establish membership in the relevant disciplinary community. The presence of citations is therefore clear evidence of dialogism and intertextuality, topics of major interest, especially since the (re)discovery of Mikhail Bakhtin in the 1980s. More instrumentally, citation counts are increasingly used to measure (sometimes perhaps injudiciously) the status and reputation of individual scholars, academic departments, institutions, and scholarly journals, thus leading to various kinds of reception study (Swales & Leeder, 2012; Paul, Charney, & Kendall, 2001; White, 2004). However, there is also convincing evidence that citations do not simply work as authorial demonstrations of due diligence with regard to previous literatures, but also operate rhetorically to strengthen arguments and claims in various ways (Gilbert, 1977; Hyland, 2004). Especially in the light of the latter, it is clear that students need not only to acquire the mechanics of citing as orchestrated by particular disciplinary conventions (APA, MLA, etc.) or to learn to avoid plagiarism, but also to embark on the arduous process of learning to cite in such a manner that their academic papers are increasingly persuasive and convincing. Indeed, much of the relevant research in English for academic purposes is ultimately designed to assist students and junior scholars, particularly nonnative speakers of English, in that process (Clugston, 2008; Davis, 2013).

This investigation makes use of a selection from the Michigan Corpus of Upper-level Student Papers (MICUSP), a freely available electronic database consisting of 829 A-graded papers grouped into 16 disciplines, totaling 2.2 million words of main text, and collected over the 2007 to 2011 period at the University of Michigan (see Römer & O'Donnell, 2011, for details). "Upper-level" indicates papers written by final-year undergraduates and graduates in their first three years of graduate study. Of the 16 disciplines, the biology subcorpus was selected for detailed citation study for several reasons. First, Ädel and Garretson (2006), in a preliminary investigation of citation patterns in an embryonic MICUSP corpus of 600,000 words, found that biology was an "outlier" and interestingly positioned somewhere between the other sciences (and engineering) and the social sciences. Hyland (2004) also found

that biology was closer to the social sciences in its frequency of citation, and Samraj (2008) showed that biology master's theses had considerably more citations than those in linguistics and philosophy. In addition, there is a rich vein of previous studies of biology writing, including Myers's wide-ranging 1990 monograph, Selzer's 1993 edited collection analyzing an article by two prominent biologists, parts of Berkenkotter and Huckin (1995), Valle's 1999 diachronic study of life science papers in the *Transactions of the Royal Society*, and a major 1994 article by Haas of particular relevance for this study. Third, according to the University website, biology at Michigan is now divided (as has been happening elsewhere) into two departments: Molecular, Cellular and Developmental Biology and Ecology and Evolutionary Biology. So, some subdisciplinary exploration should be possible. As a biology colleague once put it, "We biologists fall into two clear groups: some of us are skin-in, while others are skin-out."

In fact, over the years, citations have been categorized in various ways: in terms of their syntactic placement and linguistic form (Charles, 2006; Swales, 1990); in terms of their extent and importance (Valle, 1999); in terms of whether they are positive, neutral, tentative, or negative (Hyland, 2004); and in terms of their function or role (Harwood & Petrič, 2012; Thompson & Tribble, 2001). However, identifying the function of a particular citation is not without its difficulties. Attempting to read off citation function from purely textual evidence, as is nearly inevitable in a corpus study, is a subjective and rather chancy business (Harwood, 2008). In a recent, tightly controlled study, Willett (2013) showed that there was only "a small degree of overlap between the author's reasons for citing particular sources and their readers' subsequent perceptions of those reasons" (p. 150). Blaise Cronin (1984), arguably the doyen of citation studies, concluded in a major work that the purposes of and motives for many citations remain elusive and evasive. After all, although citation is a public act, the choice of authors to cite and the way in which they are cited may be imbued with "private intentions" (Bhatia, 2004). On the other hand, a citation may become, over time, a semiautomatic insertion as certain lines of argument become closely associated with particular people: for example, referring to genre as socially validated and regulated action will likely trigger "(Miller, 1984)," or reference to junior undergraduates moving from one disciplinary community to another as they met their distribution requirements will invoke "(Bartholomae, 1985)." Even when authors are interviewed about the forms and functions of their citing behaviors (Harwood, 2008, 2009), it sometimes turns out that no clear rationales always emerge for either form or function, even if in many cases, interesting motives can be elicited. In fact, authors interviewed about their citations ascribe more than one function in over half the instances (Harwood, 2008).

Finally, there may be mismatches between putative intentions (if any) and reader uptake. When graduate students are given (in workshops we run on academic writing) a short list of citation functions and are then asked to suggest others, they typically respond with "provide bibliographic assistance." However, it is to be doubted if article authors (excluding authors of review articles) often have this function firmly in mind. Since this is a corpus-based investigation with no access to the anonymized student authors, no attempt is made here to assign citation function.

This article *does* attempt to answer the following questions, all of which are connected to the central issue of variation in citation practice:

1. How do MICUSP students cite the previous literature in biology?
2. Do final-year undergraduates and graduate students (in their first three years) cite differently?
3. Are there noticeable differences between the citation patterns adopted in molecular versus evolutionary biology?
4. Are there differences in the citations of those using a name-and-date system as opposed to a number system?
5. What textual evidence is there of potentially evaluative engagement with previous work on the chosen biology topics?

As can be seen, the first four of these are essentially descriptive, while the last is more interpretive. The first four are disposed of relatively quickly, with the fifth being explored in greater depth.

Procedures

There are 67 texts in the entire biology subcorpus. However, the subcorpus is unbalanced in terms of level, there being 47 papers written by final-year undergraduates (coded G0), but only 9 from first-year graduates (G1), 8 from the second-year graduates (G2), and merely three from third-year graduates (G3). As a result, for the purposes of this analysis, all 20 graduate papers were examined, as were the first 26 undergraduate papers, thus achieving an approximate balance between undergraduate and graduate texts, especially since the graduate texts tend, on the whole, to be somewhat longer. In addition, a further 9 of these texts were later removed from the study because they were deemed not relevant to a study of variation in citation practice. Four contained no citations or references, being essentially short lab reports or proposed experimental designs. An additional 3 (all written by the same undergraduate) were dropped since they were summaries/reviews of a single scientific book; also discarded were 2 graduate

papers consisting of a description-cum-critique of a single journal article. The eventual working corpus, totaling just under 98,000 words, breaks down as follows: final-year undergraduates 22, first-year graduates 6, second-year graduates 7, third-year graduates 2, making 37 in all. Of course, in comparison to today's preponderance of large corpora, a corpus of around 100,000 words is very small; however, its restricted size does permit individual examination of every citation.

Research Question 1: General Aspects of the Quantitative Data

Of these 37 texts, 28 used a name and date (Harvard) system for citing, while the other 9 used a number system (Vancouver). In the whole working corpus, the average number of references per paper was 21 and the average number of citations per paper was 33. The fact that the latter number is higher might indicate that a number of references are being invoked more than once in the unfolding texts and so might suggest a fair amount of intertextual discussion and comparison. Alternatively, it might suggest that students relying on limited sources keep throwing the same references at their reader-evaluators, either to make their papers look more academic or as a way of reducing their chances of being accused of plagiarism.

In fact, the degree of citing and referencing varied widely. At the lowest end was a short (1,100 words) undergrad paper (G0.07.1) with one reference and two citations; at the top end was a long, 9,937-word third-year graduate paper, which appears to be a dissertation prospectus, with 190 references and 118 citations (G3.02.1). A more detailed breakdown of the citation frequencies (per 10,000 words) is provided in Figure 1. As can be seen, all but four of the papers fall within the middle three bands.

When we turn to citation patterns themselves, a long-standing basic distinction is the division into integral and nonintegral citations (Swales, 1990), the former placing the cited author or authors within the sentence structure (as subject or agent, or as part of a noun phrase or in an adjunct clause or phrase), while in the latter the cited author or authors occur in parenthesis. Integral citations are typically associated with a focus on the actions of researchers, while the nonintegral ones emphasize the research findings. The following pair of sentences illustrates this division:

> Hyland (2004) also employs this distinction.
>
> This distinction has been widely employed (e.g., Hyland, 2004, Thompson, 2005)

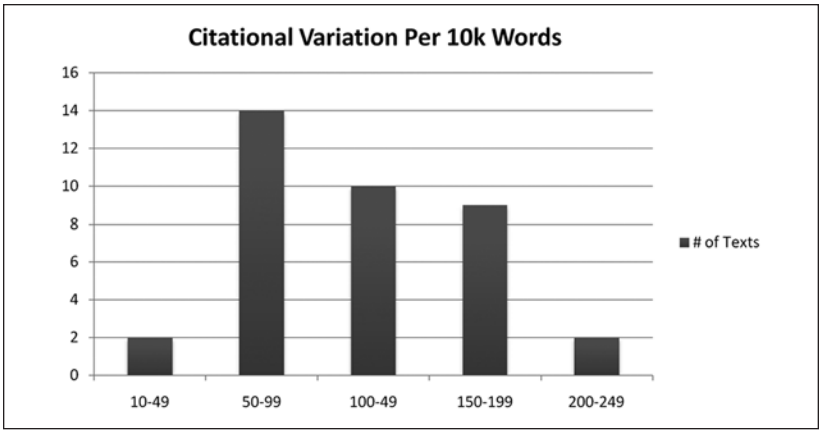

Figure 1. Citation variation per 10,000 words.

In the working corpus, there are 900 nonintegral citations and 327 integral ones, producing a percentual ratio of 73%-27%. As will be shown, this ratio is a first indication of variation in citation practice.

Finer distinctions in both categories are possible. Nonintegral citations may or may not contain a reporting verb that shows in some way what the cited author did, wrote, or thought. This distinction is illustrated as follows: Nonreporting:

> Cichlidae is a monophyletic group of perciform fishes with a species diversity approaching 2,000 described species (R2). (G2.01.1)[1]

Reporting:

> Recent phylogenetic work finds that each continental assemblage of cichlids forms its own monophyletic group with the exception of Madagascar (R7). (G2.01.1)

As might be expected, the former are much more common than the latter; of the 900 nonintegral citations, some 129 (13.7%) are reporting.

A fivefold subcategorization can be made with integral citations. The cited author or authors can function as *sentence subjects*:

> Myers (1966) hypothesized that the freshwater fishes of the West Indies dispersed from Central America. . . . (BIO.G2.01.1)

Table 1. Percentages of Integral Citation Types.

Citation type	*n*	%
Author as subject	226	70
Author as agent	26	8
Author as adjunct	19	6
Author in NP	49	15
Author (other)	7	2
Total	327	

As agent:

> It was hypothesized by Myers (1966) that the freshwater fishes of the West Indies dispersed from Central America.

As two types of adjunct:

> According to Myers (1966), freshwater fishes of the West Indies likely dispersed from Central America.

> As Myers (1966) suggests, freshwater fishes of the West Indies may have dispersed from Central America.

As can be seen, the first type is realized by a prepositional phrase, the second by a subordinate clause.

Alternatively, the author's name may *be part of a noun phrase*, either via a possessive or using an agentive structure, as in,

> Myers' 1966 hypothesis proposed that freshwater fishes. . . .

> The hypothesis proposed by Myers (1966) suggested that freshwater fishes. . . .

Finally, there is an "other" category for uses that are not common enough to merit a subcategory of their own, as in,

> In contrast to Addison et al. (1982), they argue that. . . .

The numbers and percentages for each category are shown in Table 1. As might be expected, the use of the cited author as the subject was easily the most popular category; beyond that, the table shows a clear dispersion among the other three categories. Of particular interest is the fact that the "author in

Table 2. List of Reporting Verbs.

n	Verb
48	show
31	find
23	suggest
20	propose
17	argue
11	note
10	hypothesize, do (as in "studies were done . . .")
7	demonstrate, point out
6	report, state, describe, study
5	consider, attempt to
4	analyze, conduct, assume, predict, discover
3	accept, claim, conclude, indicate, observe, reveal, test, believe, introduce, think

NP" category was easily the second most popular category—a feature worth investigating in more detail since it contrasts with certain other studies of novice writers (e.g., Mansourizadeh & Ahmed, 2011).

Another textual feature of citation practice is the choice of reporting verb. In the biology working subcorpus, a total of 112 different reporting verbs were found, again showing variation in textual practice. The more common verbs, along with their frequencies, are shown in Table 2. What is interesting about the first five is that the top two have been associated in the literature with the "hard" sciences, while the following three appear more frequently in social science and humanities texts (e.g., Hyland, 2004). It is also worth noting that *report* (typically associated with the "hard" sciences) falls outside the "top ten." Finally, a few unusual singletons might be mentioned, such as *call into question, advance the theory, espouse, echo, posit*, and *tackle the topic*.

Reporting verbs can be divided into those that are factive, that is, the writer indicates by such a choice that she or he believes that the reported proposition is correct, while nonfactive reporting verbs make no such assumption. Factives are somewhat less frequent in number and in overall frequency, although the two most common reporting verbs are factive (*show*, 48; *find*, 31). In fact, of the 31 verbs listed in Table 2, only 5 are clearly factive; in addition to *show* and *find*, we also have *demonstrate* (7), *discover* (4), and *reveal* (3). The preference for nonfactives is a further indication that many of the MICUSP biology authors are not perceiving findings from their literatures as necessarily valid, but rather are subjecting them to various kinds of intratextual reassessment.

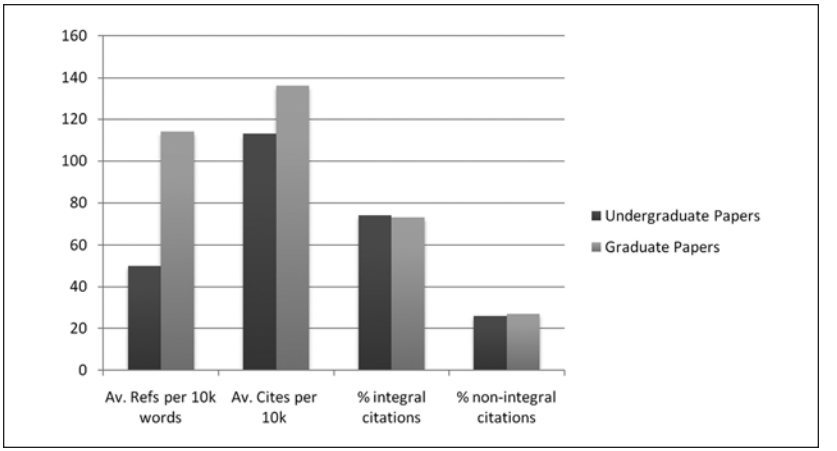

Figure 2. Undergraduate and graduate papers compared.

Research Question 2: Comparison of Undergraduate and Graduate Citational Practices

The size of the biology undergraduate corpus is 49,604 words. The average length of the 22 citing papers is 2,250 words, with an average of 10 references and 27 citations per paper. The size of graduate student biology corpus is 48,383 words, thus approximating the undergraduate corpus; however, the average length of the 15 papers increases to 3,225 words, with an average of 37 references and 45 citations per paper. Some further comparisons are shown in Figure 2.

Some of these findings need some explanation. Most obviously, the number of citations, when compared to the number of references, is more than double in the case of the undergraduate papers, while it is only about 20% larger in the graduate papers. In some respects, these figures are counterintuitive since the graduate papers are longer and contain considerably more references. Much of the answer to this conundrum lies in the fact that both subcorpora contain a number of outliers in terms of the ratio between references and citations. In fact, in all the undergraduate texts, the number of citations exceeds the number of references, but in some cases the ratio between the two is quite considerable, as in the five papers in Table 3. The last of the papers listed looks particularly anomalous; however, an explanation lies in the fact that G0.30.1 explores in considerable and comparative detail five competing hypotheses designed to explain one aspect of evolution.

Table 3. Papers With Many More Citations Than References.

Paper	# of words	# of references	# of citations	Ratio
G0.02.2	3,455	18	51	2.8:1
G0.02.5	3,173	25	73	2.9:1
G0.25.1	1,721	4	13	3.2:1
G0.01.1	5,688	12	60	5.0:1
G0.30.1	3,390	5	60	12:1

Most of the graduate papers also had more citations than references, although in two cases the numbers were almost identical. However, G3.02.1, which has already been mentioned, is the longest paper in the biology subcorpus, and also has, by some way, the most references and citations. If this paper is subtracted from the rest of the corpus, the number of references and citations per 10,000 words readjusts to 95 and 146, respectively, which is rather more in line with what might be expected when compared to the undergraduate findings. One striking similarity across the two groups is the almost identical percentages for integral and nonintegral citations. One notable difference in the references themselves is that the undergraduates make considerable use of Wikipedia and other web sites, while the graduates almost never do.

Research Question 3: Molecular and Evolutionary Biology

According to a doctoral candidate in biology, 25 of the 37 papers come from the ecological and evolutionary branch of biology (E&E), while the remaining 12 (7 undergraduate and 5 graduate) are from the molecular side. Half of the molecular papers use the number citing system, while there are only four papers in E&E biology that do not adopt the name-and-date format. The E&E papers are significantly longer (averaging just over 2,900 words as opposed to just over 2,100 in molecular) and the number of citations per 10,000 words is slightly higher (127 vs. 118). However, there are virtually no differences between percentages of nonintegral citations (73% and 74%) or between the percentages of nonintegrals employing a reporting verb (14% and 14%). In fact, this seeming parity may be misleading since one molecular paper (G0.30.1) provides over 40% of the integral citations in its subgroup and, in fact, 6 out of the 12 molecular papers contain no integral citations at all. (If G0.30.1 is removed, the nonintegral percentage rises to 81%.) Only a larger corpus would show whether these quantitative similarities are generally true of the two main branches of biology or if, more likely, the two branches diverge.

Research Question 4: The Influence of Citation System

There have been suggestions (Swales, 1990) that a number system will likely reduce the proportion of integral citations, although Charles (2006) found no clear effect in her master's of philosophy theses from politics and materials science, and neither did Clugston (2008) for health science journals. However, in the present corpus there was a clear divergence. The number-system papers had an average percentage of 88% nonintegrals, while in the name-and-date papers this percentage falls to 72%. Given that five of the nine number-system papers handle topics in molecular biology, there is presumably an interaction among citation form, citation system, and branch of biology and perhaps genre, especially with regard to "library research" papers (Nesi & Gardner, 2012), but a larger corpus would again be needed to establish this.

Research Question 5: Selected Features That Are Potentially Evaluative

Three such features have been chosen for more contextual discourse analysis: the use of a citee's first name as well as last name, the employment of direct quotations from sources, and the incorporation of source names within nominal groups. It can be argued that all three, in their strategic deployment, can be seen by readers as potentially evaluative.

First Names

A largely ignored, but potentially evaluative, feature of academic writing across both genres and types of author is the occasional use of the citees' first names, in addition to the customary last names, in integral citations. (I have personally never attested the use of a first name in a parenthetical citation.) In fact, only a single paper that covers this topic has been traced. Harwood (2008) in a Brief Communication published in an information science journal reports on his interviews with six computer scientists and six sociologists from a British university about their citation practices. In fact, first name inclusion emerged as issue only with the sociologists. One of these informants tended to include first names with integral citations, to make academic prose "less remote." Similarly, another sociologist used first names to make his papers more accessible to undergraduate readers. Several mentioned that first name usage was more typically used when citing colleagues, junior coauthors, and friends. There were also cases, such as the use of *Clifford Geertz*, where first names are more likely to be used with famous scholars.

If readers reflect on their own experience of reading academic works, they would probably conclude that in published articles and book chapters in most fields outside of the humanities, first name usage tends to be rare, except for celebratory genres such as Festschrifts, wherein the honoree may indeed be graced with his or her additional first name. With larger works, such as monographs and dissertations, first name usage may also be more common, especially in the humanities. In addition, on occasion, first names may be used simply to avoid ambiguity, and in these contexts they are definitely not evaluative. A notable instance from the field of academic discourse studies is the wife-and-husband team of Lynne and John Flowerdew.

Given the fact that this feature of academic writing has scarcely been investigated, a certain amount of speculation might not be misplaced. With that in mind, the more evaluative uses of the first name in papers would seem to fall into three main categories, although the three are not always distinguishable: (a) to acknowledge the prominence of the cited author, (b) to highlight a key reference for the citing paper's argument, or (c) to insinuate personal indebtedness to or familiarity with the cited author. On the first, the more prominent a cited author or cited work turns out to be, the increasing likelihood that the author's first name will also be used, particularly on first mention in integral format. In the second case, again first integral mention may be the likely locus. In the third case, there may be some general propensity to add first names to influential or fondly remembered mentors, or, conversely, for dissertation advisors to "boost" their ex-students.

There are three uses of first names in the graduate subcorpus. G2.02.1 is titled "Modularity and the Evolution of Complex Systems" and contains six references and 24 citations, divided equally between integral and nonintegral. On the first page occurs the one instance of a first name; it follows a sentence that explains the evolving thinking about complex systems wherein different hierarchies of organisms interact with other hierarchies in complex ways. Then we have,

> This concept, termed near decomposability by *Herbert Simon*, but more often called modularity in the biological literature, can be shown to contribute to the availability of complex systems. (my emphasis here and hereafter)

There are six further references to Simon in this paper, including a block quote, but only in the original mention is he referred to also by his first name. Of course, Herbert Simon is famous, and the use of his first name on his first appearance also aligns with his pivotal role in this paper (as shown by the seven citations to his works). The other two instances come from G2.07.1, a paper titled "Biofuels and Biodiversity." After five pages of background review, we get to the main topic, which is explored in the rest of the paper:

> The idea that perennial grasses in a polyculture can produce energy for human use is not new, as prairie-like agroecosystems have been researched by *Wes Jackson* and others at the Land Institute for many years (Jackson 2002). However, the idea that biofuels could mimic natural ecosystems was formally introduced by ecologist *David Tilman* and colleagues in 2006.

Wes Jackson is apparently well known for his work on natural systems agriculture, while David Tilman is clearly a highly influential figure. His user profile on Google Scholar (as of June 2013) yields over 70,000 citations, with a huge H-index of 124. Tilman is mentioned several times in the remainder of the paper, but his first name is not used again. The use of first names in this key section of G2.07.1 thus covertly alerts the reader to the importance of these two people for the ensuing discussion.

There are 15 undergraduate uses of first names, 13 from a single paper (G0.02.5). This is a substantial, 3,172-word library research paper titled "On the Origins of Man: Understanding the Last Two Million Years," containing 25 references and 73 citations, 45 nonintegral and 28 integral. In all cases but one, the author of this paper adopts the first name usages found in the graduate papers and in many publications. Here, for example, is paragraph eight in skeletal form:

> S1. Further, *Mark Collard's* research . . . suggests that . . .
> S2. Specifically, *Collard* proposes that . . .
> S3. Specifically, *Collard* attempts to reconstruct . . .
> S5. *Collard* attempts to make phylogenies based on . . .
> S8. However, *Collard* does not use PAUP* 4.0 for . . .
> S9. Thus, *Collard's research* may not completely invalidate the use of facial features, but it certainly casts doubt on the practical use of them. (my emphases)

As can be seen, the paragraph opens with a reference to "Mark Collard's research," which is followed by four instances of "Collard" and closes with a standard reference to "Collard's research," while at the same time, as S9 shows, concluding that Collard's findings are not in the end definitive. Once again the first name occurs in the first integral citation.

The other usage worth mentioning also occurs in the first citation of a particular work:

> Milford Wolpoff, of the University of Michigan, originally proposed this model based solely upon archeological and morphological evidence (Lahr, 1994).

This is one of those cases where the textual evidence is particularly inconclusive. Does "Milford" occur because of his prominence (he has 20 publications

with more than 100 citations), because of a pioneering role indicated by the author's use of "originally proposed," or because the author has some personal knowledge of this professor of anthropology at his own university? The remaining two first name usages also follow the established pattern. For instance, G0.1.1 has four references to Norris, three of which are parenthetical and all follow this:

> *John Norris' seminal treatise* on this question argues that there are three distinct strains of plague that have different distributions.

Overall, in practically all cases, we see first names occasionally occurring in the first integral citation to a cited author, but never later in the text. Although it might be argued that these occasional first name usages are but distant traces of having been taught MLA style during the students' freshman composition classes, their rarity, selectivity, and strategic placement would suggest otherwise. Indeed, the biology students, both graduate and undergraduate, appear to evince a fine sense of judgment in regard to this citation choice, hinting at a level of sophistication that would seem to match that of much more experienced published writers.

Direct Quotations

According to previous studies of biology research articles, direct quotation of the words of prior authors is very rare; indeed in Hyland's biology article subcorpus they were nonexistent (Hyland, 2004). The acknowledged use of others' words is also rare in the MICUSP student papers, there being just 11 examples, 3 of which merely phrasal as in "'selective neutrality' of their DNA regions (Thomson et al. 2000)" (G0.02.5). Four of the remaining eight are strategically placed, three right at the beginning and one right at the end—a position of prominence if one believes that first and last impressions can be important. Two of the other four are from G2.02.1, a discussion of Herbert Simon's theory of "near decomposability," which has already been discussed in terms of the author's use of Simon's first name. A 60-word block quote occurs on page 4, and two pages later there is another, shorter, intersentential direct quotation from the same page that the block quote was taken from. Simon thus gets the rare distinction in biology of having his own words incorporated into this text on two occasions. Two more are from G0.18.1, one of which is followed by a strongly positive evaluation:

> Li et al (1996) report that "the rate of nucleotide substitution is at least two times higher in rodents than in higher primates." This finding is particularly relevant since. . . .

The other instances are even more telling. G0.02.1 is a library research paper on sympatric speciation. It opens as follows:

> Ernst Mayr once wrote "sympatric speciation is like the Lernaean Hydra which grew two heads whenever one of its old heads was cut off." (1963:451). This observation, from his landmark text, *Animal Species and Evolution*, marks the beginning of his systematic confutation of all evidence and theory pertaining to the possibility of sympatric speciation available at that time. Much has changed, however, in the field of biology since '63, but his point still remains valid: . . .

According to Wikipedia, Ernst Mayr was "one of the 20th century's leading evolutionary biologists," and here he comes across as receiving a triple, albeit covert, accolade: He opens the story, his first name is used, and he is granted the opportunity to speak in his own words. There are other noteworthy aspects of this undergraduate paper, and this despite the rather awkward second half of the second sentence. In particular, the author appears highly self-confident in his appraisal of the development of his field: A "landmark text . . . marks the beginning" and "Much has changed . . . since '63" show a firm historical sense, not to speak of the employment of the really rare noun *confutation*. (There are but 12 examples out of 450 million words in COCA.)

The other opening involving direct quotation comes from G1.04.1, which is titled "The Evolution of Terrestriality: A Look at the Factors That Drove Tetrapods to Move Onto Land." The first two sentences read,

> The fish-tetrapod transition has been called "the greatest step in vertebrate history" (Long and Gordon, 2004) and even "one of the most significant events in the history of life" (Carroll, 2001). Indeed, the morphological, physiological and behavioral changes necessary for such a transformation in lifestyle to occur are astounding.

The two quotations thus anchor the paper, with the two quoted superlative phrases underscoring for the reader the importance of the chosen topic. The remaining extensive direct quotation comes from G0.02.5 on the origins of man, which we have already met in connection with its frequent use of first names, but in this case it actually closes the paper:

> Recent attempts to make models that are more complex have also fallen short (Eswaran, 2002; Excoffier, 2002). Erik Trinkaus, of Washington University, best summons up the current state of human origin theory:
>
> > "I believe that it [Eswaran's model] suffers from a problem shared with the majority of the current and past models of modern human emergence:

> namely, it tries to explain too much of a geographically and temporally complex process with a single mechanism . . .” (Eswaran, 2002, 767).

Another interesting mini-text! It turns out that Erik Trinkaus is a very well-known authority on early man and that this quotation, which is not cited among the 25 references, actually comes from his printed commentary on Eswaran’s article. As for “best summons up,” it is not clear whether that is simple slip not detectable by the spell-checker for “best sums up,” or is a clever play on words. So it emerges that the actual words of other scholars, on the whole famous ones, tend to be placed as either the first words or the last words. Given the rarity of invoking the actual words of others in biology, the choice of doing so again appears to reflect, like first naming, a selective rhetorical strategy on the part of MICUSP contributors.

Author Names in NP Structures

Lancaster (2012) investigated differences between high-performing and low-performing written assignments in senior undergraduate courses in economics and political science. One of the differences that emerged from his study was that the high-performing writers had a greater tendency to use “concept-focused” rather than “person-focused” citations. By the former, Lancaster means uses such as “Rawls’ argument implies that . . .” and the latter “Rawls argues. . . .” The “concept-focused” citations, it can be argued, tend to show greater conceptual integration of the cited sources. And, at this juncture, it may be recalled that one of the small surprises in the breakdown of subtypes of integral citations was the fact that “author in NP” was the second most common category, composing 15% of the total. In fact, there are four graduate papers with four or more NP citations. These four papers (G2.07.1, G1.05.1, G2.01.1, and G1.04.1) are papers that have been mentioned before since they are all examples of rich intertextual storytelling. All are quite long, around 4,000 words of main text, and all have numerous citations (in all cases over 50), with relatively large minorities of integral citations. Apart from one of two procedural nouns, the rest are clearly conceptual, as evidenced by the nouns chosen: *results* (3), *hypothesis* (2), *analysis* (2), and the following singletons—*objections, emphasis, focus, concept, model, approach, picture, metaphor*, and *view*.

The following three examples, all from G2.02.1, illustrate the way this author engages with the literature:

> Leon Croizat’s (1962) metaphor of vicariance biogeography being like reconstructing a pane of glass that has been repeatedly shattered seems particularly relevant to the Greater Antilles.

> In the alternative view proposed by Iturralde-Vinent and MacPhee (1999) based on geological evidence, a short-lived connection between the Greater Antillean Islands . . . and northwest South America existed circa 32 million years ago.

> The absence of cichlids from Jamaica and particularly Puerto Rico does not bode well for the Iturralde-Vinent and MacPhee (1999) hypothesis.

And finally, a double example from G1.04.1:

> This recalls Ewer's (1955) emphasis on the importance of population pressure, as well as Goin and Goin's (1956) focus on competition in tetrapod evolution.

All in all, this subgroup of graduate papers demonstrates considerable variation in citational patterning, with rare but rhetorically marked use of the three alternate usages discussed in this section.

Only three of the undergraduate papers had three or more instances of "concept-focused" NP citations (G0.02.5, G0.18.1, and G0.30.1). On the whole, they are more straightforward than the graduate uses, at least in the sense that most of the examples simply iterate or reiterate a model or hypothesis associated with a particular scholar or group of scholars. Here, as an illustration, are the final two sentences from the introductory section of G0.30.1, which is titled "Assessing Selection Hypotheses for the CCR5-Δ32 Mutation in Europeans":

> This paper seeks to investigate the cause(s) of positive selection proposed by these hypotheses and models. The hypotheses and models of interest include: Duncan et al's plague hypothesis, Galvani and Slatkin's smallpox hypothesis, Balanovsky et al's ecological factors model, and a Bronze Age hypothesis suggested by Sabeti et al and Hedrick and Verrelli.

In all, *model* occurred 5 times in the undergraduate subcorpus, *hypothesis* 4, and *modeling* 3, with single occurrences of *contention, research, results*, and *method*.

Variations in Citational Patterning

So far, the focus has been on those papers that offer textual variety in their accounting for attributed previous work on their topics, be it via a balance of nonintegral and integral citations, via a range of integral subtypes, by the occasional uses of direct quotation or author first names, or by using author names in NP structures. Most of these papers are, not unexpectedly, library

research papers, and most come from the evolutionary or ecological side of biology. The contrasting nonreporting nonintegral citing style occurs quite widely in many published papers and is typically used in the working corpus among nearly all the student writers when covering background material subsidiary to the main arguments. However, it was also used in a minority of papers (12 undergraduate and 3 graduate) as the only or almost only vehicle for acknowledging the work of others. Take the case, for example, of a long (over 5,500 words) female undergraduate paper titled "The Ecology and Epidemiology of Plague." This contains 60 citations, of which all but 2 are parenthetical, and of these 58 only 4 contain a reporting verb. Here is part of the paragraph that opens the description of the Plague bacterium:

> This bacterium is a non-motile non-sporeforming, Gram-negative coccobacillus (Bahmanyar and Cavanaugh, 1976). *Y. pestis* is very fragile, and can be killed by heat-treatment at 55°C, chemical agents, sunlight, or extreme dryness (Stark *et al*., 1966). This is why the bacterium is generally not found in extremely arid environments, such as the Saharan Desert and the Middle East. Its optimum growth temperature is 28°C, so it prefers subtropical climates, but it has been found to grow in full nutrient broth at temperatures ranging from −2 degrees C to 45°C (Stark *et al*., 1966).[2] *In vivo*, where nutrients are available, its optimal growth temperature is 37°C, the body temperature of mammals (Rail, 1985). (G0.01.1, p. 9)

There is a certain repetitive quality to this consistent citing style that in the title I referred to—perhaps unfairly—as "parenthetical plonking." This practice is sometimes caricatured as "nods all round to previous researchers." Although, as the above paragraph shows, the author of the Plague text is a competent academic writer, especially given her place on the educational ladder, this lack of variety in citation patterning contrasts with many of the other papers in the corpus.

As a counterpoint to G0.01.1, we can finally take the case of G0.02.2, a 3,500-word undergraduate paper titled "Host-Parasite Interactions: On the Presumed Sympatric Speciation in *Vidua*." Here are the opening sentences from a section headed "Evidence for Divergence in Sympatry":

> Theoretical, verbal, and mathematical models can only go so far—descriptive and exhaustive field research is essential in showing what is actually happening. To this end, Sorenson and Payne have devoted a large portion of their fieldwork to genetics (Sorenson & Payne, 2001, 2002; Sorenson *et al*. 2003, 2004). Their research efforts have focused on the creation of an accurate phylogeny based on genetic data—obvious arguments against this methodology and its application have been raised by many (see Coyne & Orr, 2004); however, their conclusions seem valid (Sorenson *et al*. 2003; Sefc *et al*. 2005).

The section opens with a highly evaluative statement about field research being "essential." This is linked by "To this end," which is followed by the two principal players in this paper, who are placed in subject position. The opening of the third sentence follows up with "Their research efforts," goes on to discuss possible problems with the methodology, but ends with the suggestion that it probably works. Also note the use of "see" in the Coyne and Orr citation. Without the "see," this would be what is sometimes described as an ambiguous "hanging" citation (Swales & Feak, 2004): Did Coyne and Orr raise the objections, or did they report them? In fact, the use of "see" strongly implies that their work reviews the arguments against the methodology. More generally, this extract is quite typical of the rest of this paper; it evaluates the science on its topic, problematizing where the author thinks appropriate (Barton, 2002), and discusses work on the topic with exemplary variation. There are 51 citations in the paper, 23 parenthetical and 28 integral; of the latter, 22 use author as subject, 4 as passive agent, and 2 as part of a noun phrase. There are also two direct quotations, and the first time Payne and Sorenson are introduced their first names are added. If one wanted an undergraduate paper that showed intertextual and intratextual complexity and variability in its citation practice, G0.02.2 would be a strong candidate.

Discussion

This study has shown considerable variation in citation practice—in its various manifestations as discussed above—in a large majority of the papers in the working MICUSP biology subcorpus. Of course, we do not know the factors that might have led to this variation. The corpus provides no information about assignment instructions, such as required length, citation system to be used, number of references expected, or indeed whether the student writers could have taken advantage of instructor or peer reviewer comments on preliminary drafts; as a result, these variables cannot be factored in. Another issue that cannot be addressed is the possible effect that the frequency and type of citation might have had on the grade since the corpus contains no lower-graded papers for comparison purposes (unlike in Lancaster's 2012 dissertation).

In her longitudinal study, Haas (1994) shows how "Eliza," as she proceeded toward her degree in biology, had, by the time of her fourth and final year, learned to see her biology readings not as the faceless products of scientific investigations, but as being created by "authors." As Haas says, "Eliza's attention to rhetorical elements of discourse—authors, readers, motives, contexts—also exhibited increased sophistication in her senior year" (p. 66). Given that higher-level academic reading skills are a necessary if not

sufficient condition for higher-level academic writing skills (Shaw & Pecorari, 2012), then it would seem, or so the textual evidence would intimate, that a majority of the MICUSP biology writers in the working corpus also (like Eliza) perceive their references as being "authored" by individuals or groups engaged in negotiating various kinds of knowledge claim (Myers, 1990) or in putting forward various kinds of proposal.

Indeed, from a citation perspective, the biology subcorpus is denser and more variously patterned than most of the other 15 disciplines collected in MICUSP. Preliminary examination of papers in civil engineering, economics, education, linguistics, and nursing showed fewer and less evaluative citations than in biology. Ädel and Garretson (2006) found, in terms of "other reference,"[3] that its occurrence in biology per 10,000 words was exceeded only by philosophy and sociology. However, the working corpus nonintegral/integral split of 77% to 23% differs from that found by Samraj (2008), with her split of 88% to 12% for biology master's theses, and differs from Hyland's finding of 90% to 10% for published research articles in biology, plus the fact that there were zero occurrences of direct quotation in the biology articles (Hyland, 2004). Ädel and Garretson (2006) observe, "[T]he fact that Hyland consistently has more non-integral structures than MICUSP could be due to the editing process and size restrictions of academic journals, but it probably indicates a strong learning curve in the use of non-integral forms" (p. 278). Mansourizadeh and Ahmed (2011) take a similar position when discussing differences between novice and expert Malay chemical engineers writing in English. The experts' ratio was 86% to 14% nonintegral/integral, while for the novices it was 73% to 27%. They also note,

> In expert writers' papers, there were almost equal quantities of integral-verb controlling [reporting] and naming citations, while in the novices' papers, the verb-controlling citations were used five times more than naming citations. This could be due to the novices' lack of skill in constructing nominalization and complex noun phrases, both of which typically pose problems for beginning writers. (p. 157)

Finally, it should be noted that Samraj ascribes her high percentage of parentheticals to the considerable use made of generalizations (i.e., here several sources are cited together).

A number of comments seem in order here. First, Hyland's biologists all worked in the molecular branch of the field, which tends to be more experimental, "harder," and less heavily contextualized than the E&E branch. Second, it is undoubtedly true that when students first learn to use references in their academic writing (probably in the last years of high school), they will

be using integral forms, as in "As Shakespeare says, 'troubles come not as single spies.'" In addition, they will most likely have had practice in the mechanics of citing during freshman composition classes. But that said, the students in this study had at least 3 years of exposure to academic writing at a major research university, and, in consequence, it could be argued that they could have used more parenthetical citations, but they chose not to do so because, in most of the cases illustrated in this paper, they are interacting in an overtly cognizant and intertextual manner with readings relevant to their topics. And it is interesting that Lancaster (2012) reached similar conclusions in his study of high-performing papers written by students in upper-level intensive writing classes in largely nonquantitative economics and political science. Finally, the MICUSP data analyzed in this study produce a lower proportion of generalizations than in Samraj (2008), as indeed suggested by the undergraduate usage shown in Table 3.

In some contrast, Charles's results for master's of philosophy theses in politics and materials science show integral citation frequencies of 47% and 55%, respectively, and Thompson and Tribble give percentages of 33.5% and 62% for agricultural botany and agricultural economics doctoral dissertations (Charles, 2006; Thompson & Tribble, 2001). As Charles (2006) notes, it is probable that the length of these genre exemplars is the most likely explanation. If a citation is extensive (i.e., covering several sentences or a paragraph, as in the "Mark Collard" extract shown in the First Names subsection above), then there is a tendency to rely on authorial subjects or their pronominal equivalents as a way of managing stretches of descriptive or evaluative detail. In sum, in a collection of scientific papers averaging under 3,000 words, an integral proportion of around a quarter would seem to indicate an intellectual engagement with the literature rather than a regularized acceptance of the "facts" reported in that literature.

There are criticisms in the literature (Harwood, 2009, 2010; Thompson & Tribble, 2001) about the quality of teaching materials on citation, particularly because of the stress on the mechanics of citing, rather than on its wider and more rhetorical role in orchestrating academic contexts and arguments. However, the variation in citation practice shown in many of the "A" papers suggests that their authors are cognizant of the citation choices available and, more speculatively, of their possible effects on the reader. So, in terms of the selected features of academic writing discussed in this study, it would seem "all's well that ends well." More broadly, this study has attempted to show what can be done by way of detailed analysis with a freely and easily available collection of student writing (or subset thereof). In fact, the MICUSP corpus was originally constructed with one main aim and one subsidiary one. The main aim was to provide examples of student disciplinary writing

situated between the freshman year (of which much is known) and the other end, consisting of dissertations and published scholarship (on which many studies have been conducted). The subsidiary aim was to provide evidence for the timing of Cheryl Geisler's great divide model of academic literacy as to when students move from being consumers of knowledge to interact with it (Geisler, 1994)—a move often associated with fourth-year undergraduates or the first years in graduate school. But, as it turns out and as this study shows, the MICUSP undergraduate biology students are "just too damn good."

Acknowledgments

I would like to thank the three anonymous reviewers for many helpful suggestions. Many thanks also to Chris Walczak for subdividing the biology papers and to my part-time research assistant, Kohlee Kennedy.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The R plus number indicates the number of references listed.
2. It could be argued that "has been found to grow" is reporting, but I took it as stative, as in "Coal is found in the ground" (Svartvik, 1985).
3. "Other reference" is a somewhat broader category than the "citation" criterion used in this study since it includes historical nominations such as "Le Pen received almost 17% of the vote."

References

Ädel, A., & Garretson, G. (2006). Citation practices across the disciplines: The case of proficient student writing. In C. Perez-Llantada, R. Plo, & C.-P. Neumann (Eds.), *Academic and professional communication in the 21st century: Genres and rhetoric in the construction of disciplinary knowledge* (pp. 271-280). Zaragoza, Spain: Prensas Universitarias.

Barton, E. (2002). Inductive discourse analysis: Discovering rich features. In E. Barton & G. Stygall (Eds.), *Discourse studies in composition* (pp. 19-42). Cresskill, NJ: Hampton Press.

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

Bhatia, V. K. (2004). *Worlds of written discourse: A genre-based view*. London, UK: Continuum.

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*, 310-331.

Clugston, M. (2008). An analysis of citation forms in health science journals. *Journal of Academic Language and Learning*, *2*, 11-22.

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

Davis, M. (2013). The development of source use by international postgraduate students. *Journal of English for Academic Purposes*, *12*, 125-135.

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

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

Haas, C. (1994). Learning to read biology: One student's rhetorical development in college. *Written Communication*, *11*, 43-84.

Harwood, N. (2008). Citers' use of citees' names: Findings from a qualitative interview-based study. *Journal of the American Society for Information Science and Technology*, *59*, 1007-1011.

Harwood, N. (2009). An interview-based study of the functions of citations in academic writing across two disciplines. *Journal of Pragmatics*, *41*, 497-518.

Harwood, N. (2010). Research-based materials to demystify academic citation for postgraduates. In N. Harwood (Ed.), *ELT materials development* (pp. 306-326). Cambridge, UK: Cambridge University Press.

Harwood, N., & Petrič, B. (2012). Performance in the citing behavior of two student writers. *Written Communication*, *29*, 55-103.

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

Lancaster, Z. (2012). *Stance and reader positioning in upper-level student writing in political theory and economics* (Unpublished doctoral dissertation). University of Michigan, Ann Arbor.

Mansourizadeh, K., & Ahmed, U. K. (2011). Citation practices among non-native expert and novice scientific writers. *Journal of English for Academic Purposes*, *10*, 152-161.

Myers, G. (1990). *Writing biology: Texts in the social construction of scientific knowledge*. Madison: University of Wisconsin Press.

Nesi, H., & Gardner, S. (2102). *Genres across the disciplines: Student writing in higher education*. Cambridge, UK: Cambridge University Press.

Paul, D., Charney, D., & Kendall, A. (2001). Moving beyond the moment: Reception studies in the rhetoric of science. *Journal of Technical and Business Communication*, *15*, 372-399.

Römer, U., & O'Donnell, M. (2011). From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, *6*, 159-177.

Samraj, B. (2008). A discourse analysis of master's theses across disciplines with a focus on introductions. *Journal of English for Academic Purposes*, *7*, 55-67.

Selzer, J. (Ed.). (1993). *Understanding scientific prose*. Madison: University of Wisconsin Press.

Shaw, P., & Pecorari, D. (2012). Source use in academic writing: An introduction to the special issue. *Journal of English for Academic Purposes*, *11*, 1-3.

Svartvik, J. (1985). *On voice in the English verb*. The Hague, Netherlands: Mouton.

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

Swales, J. M., & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills*. Ann Arbor: University of Michigan Press.

Swales, J. M., & Leeder, C. (2012). A reception study of articles published in *English for Specific Purposes* from 1990 to 1999. *English for Specific Purposes, 31*, 137-146.

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

Valle, E. (1999). *A collective intelligence: The life sciences in the Royal Society as a scientific discourse community, 1665-1965* (Anglicana Turkuensia No. 17). Turku, Finland: University of Turku.

White, H. D. (2004). Citation analysis and discourse analysis revisited. *Applied Linguistics*, *25*, 89-116.

Willett, P. (2013). Readers' perceptions of authors' citing behavior. *Journal of Documentation*, *69*, 145-156.

Author Biography

John M. Swales is professor emeritus of linguistics at the University of Michigan, where he was director of the English Language Institute from 1985 to 2001. Recent book-length publications include *Incidents in an Educational Life: A Memoir of Sorts* (2009), *Aspects of Article Introductions* (2011; a reissue, with a new introduction, of the 1981 monograph), and the third edition (with Christine Feak) of *Academic Writing for Graduate Students* (2012), all published by the University of Michigan Press.