Citation practices across the disciplines:
The case of proficient student writing

Annelie Ådel
University of Michigan
Gregory Garretson
Boston University

Introduction
A central feature of academic writing is references to other sources, as has been
demonstrated in studies by Hyland (1999) and Tadros (1993). While this research has
emphasized the importance of attribution in published academic writing—especially as
it illustrates a tension between originality and humility to the community (Myers, 1990;
Berkenkotter & Huckin, 1995)—attribution in unpublished academic writing by
advanced students is uncharted territory. Here we present a first foray into this territory:
a pilot study in which we analyzed other-reference in a 600,000-word set of papers from
the MICUSP corpus.

MICUSP, the Michigan Corpus of Upper-level Student Papers, is a new corpus of
student writing in English. Under development at the University of Michigan’s English
Language Institute, MICUSP consists of A-grade writing by senior undergraduate and
graduate students in the US. The corpus will eventually be fully balanced with respect
to the ten different disciplines included (see Figure 5), year of study of the writers
(ranging from fourth-year undergraduate to third-year graduate), and possibly also with
respect to native vs. non-native speaker status. The gap that we intend MICUSP to fill
has been described by Nesi et al. (2004:440) as follows: “Far more academic writing is
produced for assessment purposes than for publication purposes, but because of the lack
of a suitable corpus, research into the generic features of published academic writing
vastly outweighs research into the generic features of assessed student writing”. Like
MICASE (Simpson et al., 2002), this corpus will be made available to the research
community upon completion.

Research questions
The starting point of this research is a 1999 study by Ken Hyland on attribution in
research articles from eight different disciplines; this identified clear disciplinary
differences in the extent to which writers refer to the work of others. Our overall
research question is how advanced student writers perform other-reference. Focussing
on inter-disciplinary variation, we ask these specific questions: (i) how do these student
writers refer to the work of others, (ii) how are these references realized, and (iii) are the
same disciplinary differences that have been found in professional writing (as shown by
Hyland, 1999) evident in student writing? In order to compare student and professional
writing, we will refer to results from the 500,000-word corpus described in Hyland
(1999) (see figure 5).
Types of other-reference

Hyland (1999:341) refers to “reference to prior research” as “the attribution of propositional content to another source”. In practice, this is quite a broad definition, including instances not only like those in (1) and (2), but also like those in (3).

(1)  a. It is the individual’s given expressions that define the basis for how the group will interact with the individual (Goffman, 1960). [SOC]
    b. The discovery of Frasnian-age panderichthyids and tetrapods in Latvia and Russia offer strong support for a Euramerican origin of tetrapods (Ahlberg, 1995; Clack 2002b; Daeschler and Shubin, 1995). [BIO]

(2)  a. Bourdieu argues that the subjective moment is crucial in constructing class since “agents are both classified and classifiers” (Bourdieu 1987: 5). [SOC]
    b. In this way, he suggests, efficiencies can be made at the intramodular level without disturbing the balance of the system. [BIO]

(3)  a. Durkheim’s theory is very attractive [SOC]
    b. Le Pen received almost 17 percent of the vote - only 3 percent less than Chirac [SOC]
    c. Based on the positions of Barnet and Tucker briefly outlined above, one can conclude that the writer constructs an interpretation of an artwork based on both evaluation […] and description. [LIN]

We make a distinction between citation (1), a non-integral reference to a source, and attribution (2), the act of ascribing some (linguistic or cognitive) act to a source in a syntactically integrated way.\(^1\) Furthermore, we make a distinction between both of these and mere mention (3). Simply mentioning a source’s name does not automatically make for an act of attribution, especially if the source is not a semantic agent (illustrated in 3a-b). Other researchers and texts are, as one might expect, very often mentioned in academic writing. However, many such references play a general referential or cohesive role in the discourse (illustrated in 3c) and do not constitute attribution. Our umbrella term for these three types is other-reference.

Method

Our study focussed on other-reference in the form of (i) third person pronouns, (ii) proper names, (iii) a selection of nouns commonly used in attribution (e.g. researcher and author), and (iv) numerical citation forms. We extracted all instances of (i)-(iv) from the MICUSP pilot corpus using WordSmith Tools (Scott, 1999), and then went through all instances, carefully coding them for a range of features (most of which we will report on below), with the aid of a custom-made coding program called the Attribution Coder.\(^2\) We then retrieved the number of occurrences and distribution of all the features under analysis by means of Perl scripts.

Results

The ten disciplines included in the corpus, and their abbreviations, are the following: philosophy (PHI), sociology (SOC), biology (BIO), psychology (PSY), nursing (NUR),
civil engineering (CEE), linguistics (LIN), physics (PHY), industrial and operations engineering (IOE), and economics (ECO). The physics subsection is very small, so we must be highly tentative in drawing conclusions about that discipline.

Figure 1 illustrates the extent to which writers refer to the work of others in the different MICUSP disciplines.

Figure 1. Amount other-reference by discipline

We get the expected differentiation between “hard” and “soft” disciplines only to some degree. Philosophy, sociology, and psychology are at the top, while the engineering fields are toward the bottom. Linguistics, however, is over at the lower end, while biology, a traditionally hard discipline, has almost the same frequency as sociology. Perhaps the most surprising aspect of the distribution is the considerable variation in the amount of other-reference: economics has only one-fourth as much as philosophy, yielding a range of 45–201 tokens per 10,000 words.

Figure 2 shows how integral forms (represented by Attribution and Mention) and non-integral forms (represented by Citation) are distributed across the disciplines. As we might expect, attribution is more common in the “soft” disciplines, while citation is clearly preferred in the hard sciences. Psychology and IO Engineering are the only disciplines not following this pattern. In general, in the humanities and social sciences, writers not only frequently attribute statements and acts to other researchers (Attribution), but also spend considerable time discussing fellow researchers and their ideas, as well as “classic” theories (Mention).
Figure 2. Types of other-reference by discipline

Figure 2: Types of other-reference by discipline

Figure 3 shows the distribution of verbal, nominal and other types of attribution.

Figure 3. Distribution of verbal, nominal and other types of attribution by discipline

The typical attribution type is a verbal structure, involving a reporting verb, as in (2) above. Nominal types involve reporting nouns, as in (4) below, while the remaining category predominantly involves prepositions, as in (5).

(4) it is difficult to reconcile Panksepp’s notion of 50s as indicators of general positive affect or joy with... [PSY]
a. According to Pitman et al. (1993), Cuba and Hispaniola did not separate until a shearing in the late Middle Eocene. [BIO]

b. For Butler the paradox of redefinition and historicity resides in the performative… [SOC]

Overall, verbal forms are vastly preferred. There is a very slight tendency for the soft disciplines to use nominal types, which perhaps suggests a greater stylistic preference for nominalizations in those disciplines.

Table 1 shows the most common verbs used in cases of attribution only. Some simple and complex verbs have been collapsed in the table; make statement, for example, was grouped with state, draw/make conclusion with conclude, etc. Physics is excluded, as it contains too few texts to provide any reliable numbers.

<table>
<thead>
<tr>
<th>Disc.</th>
<th>Top-ranking verbs (starting with most frequent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIO</td>
<td>note, suggest, propose, argue, find, cite, describe, observe, explore, show</td>
</tr>
<tr>
<td>CEE</td>
<td>cite, note, find, argue, determine, focus on, contradict, define, discover, propose</td>
</tr>
<tr>
<td>ECO</td>
<td>find, show, argue, identify, point out, propose, present</td>
</tr>
<tr>
<td>IOE</td>
<td>suggest, use, find, say, conclude, mention, study, state, investigate, propose</td>
</tr>
<tr>
<td>LIN</td>
<td>argue, say, claim, point out, describe, discuss, suggest, state, propose, mention, focus on</td>
</tr>
<tr>
<td>NUR</td>
<td>state, discuss, tell, say, describe, recommend, suggest, stress, talk</td>
</tr>
<tr>
<td>PHI</td>
<td>argue, claim, say, note, state, think, suggest, believe, mean, conclude</td>
</tr>
<tr>
<td>PSY</td>
<td>find, suggest, argue, use, examine, point out, discuss, show, state, conclude</td>
</tr>
<tr>
<td>SOC</td>
<td>argue, define, call, state, see, believe, say, write, describe, suggest</td>
</tr>
</tbody>
</table>

There is great variation in the use of reporting verbs in MICUSP. As many as 15 of the 88 verbs listed occur on the list for only one discipline. The “soft” verb point out is found only in psychology, linguistics, and economics. Argue is important in sociology and philosophy and is the most common verb in linguistics. Claim is on the philosophy and linguistics lists only. Think and mean are unique to the philosophy list. Overall, the humanities and social sciences have a larger number of prototypical verba dicendi than the hard disciplines. In the hard sciences, the verbs observe and explore make it onto the biology list, while investigate makes it onto the IO Engineering list. Overall, however, reporting verbs seem to used be rather idiosyncratically in each discipline, and it is hard to find clear patterns across the hard-soft discipline divide. A large number of the verbs, like show (biology, economics, psychology) and describe (biology, linguistics, nursing, sociology), pattern in seemingly random ways.

Figure 4 shows what we term faithfulness in cases of attribution: a record of whether reported units are quoted verbatim or are rephrased.
With faithfulness in focus, we see a more traditional divide between hard and soft disciplines. Looking at the data this way, biology ends up at the low-frequency end (cf. Figure 1). Linguistics, philosophy and sociology now cluster at the top.

**Comparison with professional writing**

We now turn to the question of how the results of our MICUSP study compare to those from Hyland’s (1999) study of published research articles. Figure 5 illustrates the differences between the corpora, and the extent to which the disciplines may be compared. The four engineering disciplines in the two corpora are arguably quite distinct; we must therefore proceed with caution when comparing these disciplines.
There are no major differences between the student and professional writers, except in philosophy and sociology. Because other-reference is such a salient feature of these
fields, perhaps it is easier for students to overdo it when trying to model “philosophy writing” and “sociology writing”? The results also confirm one of Hyland’s findings—that biology stands out among the sciences. The exceptional frequencies found in Hyland are matched in MICUSP.

Figure 7 illustrates the distribution of integral vs. non-integral types in MICUSP and Hyland.

Figure 7. Distribution of integral vs. non-integral types

Students and professional writers use these forms very differently. In some cases, the pattern is the same while the frequencies are different (philosophy, biology, engineering), but in others, the two types pattern in exactly opposite ways (sociology, linguistics, physics). The 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 also indicates a steep learning curve in the use of non-integral forms.

It is not possible to compare verbal, nominal and other types of attribution in the two corpora, since Hyland does not consider forms other than verbal ones. We can, however, take a closer look at the verbs. Figure 8 shows the most common reporting verbs for each of the comparable disciplines, and the cases of overlap.
Figure 8. Most common reporting verbs in MICUSP and Hyland corpus

With the exception of philosophy, there is little overlap between the lexical preferences of the students and the professionals. This demonstrates that disciplines—even subdisciplines, as in the case of the engineering fields—each tend to speak their own “dialect”. In engineering, only one out of the 13 verb types listed (use) is shared by both corpora; and even within the corpora, there are few verbs shared between disciplines.

While we might expect student writers to use a small range of non-discipline-specific reporting verbs, the MICUSP data clearly show that they actually use a wide variety of relatively discipline-specific verbs. The variation is much greater than in professional writing: Hyland found over 400 different verbs in his 500,000-word corpus, while we found over 700 different verbs in our 600,000-word corpus. The curious aspect is that, as we have seen, the verbs favoured by the students are not those favoured by the more experienced academics.

Without access to Hyland’s original data, we were unable to compare faithfulness in the two corpora in detail. What we can say, however, is that verbatim quoting in Hyland was minimal and it “did not occur in any science papers” (Hyland, 1999: 348). Where it did occur, sociology had the most quotes (13%), followed by linguistics (10%) and philosophy (2%). Comparing these numbers to those in Figure 4, we see that linguistics exceeds 20%, followed by philosophy and sociology at 19%. With the exception of physics, quotes occur even in the hard sciences in MICUSP.
The frequent use of verbatim quotes could be the aspect that most marks the MICUSP writers as students; they appear, proportionally, to give more weight to the words of their authoritative sources.

Hyland (1999:341) summarizes his findings by saying that writers in the humanities and social sciences (i) employed substantially more citations than scientists and engineers and (ii) were more likely to use integral structures. Generally, these findings hold true for MICUSP, but with biology and linguistics as “outliers” with respect to (i), and psychology and IOE with respect to (ii). Like Hyland’s, our results also support a distinction between “the persuasive purposes of academic citation in different traditions” (Hyland, 1999: 348). We can conclude that these student writers already appear to be highly specialized and well socialized into their respective disciplines.

**Future directions**

Much remains to be done to analyze further this rich set of data. It remains to be determined to what degree genre influences our results. While MICUSP consists of different genres (mainly term papers, but also case studies, literature reviews, lab reports and reaction papers), the Hyland corpus consists exclusively of research articles. We also plan to look for possible differences in other-reference between the small non-native-speaker portion (20%) and the native-speaker portion (80%) of MICUSP.

An additional research question posed in this project concerns how writers depict the information they report. We hypothesize that stance is unusually integrated with attribution in student writing. Our corpus has been coded for stance, and we plan to analyze these data in the near future.

**References**


---

1 The division of citation forms into ‘integral’ and ‘non-integral’ is Swales’ (1990). This is a distinction between citations that are syntactically integrated into the text and those that are syntactically non-integrated (usually separated by paragraphs or brackets).
2 We would like to acknowledge the help from Rebecca Maybaum in coding a subset of the examples.
3 Due to space limitations, we will not discuss nouns here.
4 We also encountered a small number of a “hybrid” type in MICUSP, as in *(Bayer et al. 2005)*’s System 2 is exactly this type of system. [LIN].