Attended/unattended *this* in academic student writing: Quantitative and qualitative perspectives*

STEFANIE WULFF, UTE RÖMER and JOHN SWALES

Abstract

This paper addresses the question of what governs the optional attendance of the determiner this *by a noun phrase in academic student writing. Previous research on* this *has largely focused on the noun phrases accompanying* this*, while the question of what determines writers' choice between attended and unattended* this *in the first place has received only little attention. In the present study, we present the results of a more comprehensive analysis, including quantitative methods (logistic regression analysis, Distinctive Collexeme Analysis, textual distribution measures) and qualitative methods (cluster extraction), of more than 5,800 hits of sentence-initial* this *obtained from the* Michigan Corpus of Upper Level Student Papers *(MICUSP). Overall, the results point to a strong influence of the verb accompanying (un)attended* this*, which is moderated to some extent by author-related variables like academic discipline, academic proficiency level, native speaker status, and gender. A qualitative pattern analysis of the most prominent* this *+ verb clusters reveals that semantic biases evidenced in the verbs distinctively associated with (un)attended* this *are reflected at the text-organizational level in terms of positional preferences within paragraphs and texts. In combination, the results point towards an ongoing delexicalization of* this *+ verb clusters like* this is *and* this means *into textual organization markers, which stands in sharp contrast to traditional cautions against unattended* this *as mere "vague reference" that is to be avoided.*

Keywords: *disciplinary variation, distinctive collexeme analysis, clusters, logistic regression, multifactorial analysis, phraseology, student academic writing, textual distribution|organization, (un)attended* this

Corpus Linguistics and Linguistic Theory 8–1 (2012), 129–157 1613-7027/12/0008–0129
DOI 10.1515/cllt-2012-0006

1. Introduction

The extremely common English demonstrative form *this* can function either as a free-standing pronoun as illustrated in (1) or as a determiner attending a head noun phrase as in (2).

(1) *This* is an example.
(2) *This* sentence is an example.

As we review in Section 2 below, previous research on *this* has largely been restricted to analyses of the particular noun phrases that attend *this* (e.g., Francis 1986; Charles 2003), mostly to disambiguate antecedents in the previous text. The question what determines writers' choice between unattended and attended *this* in the first place has, on the contrary, received only little attention. The present study seeks to take a first step towards closing this gap by considering instances of sentence-initial *this* in academic student writing. We have based our analysis on more than 5,800 hits obtained from a preliminary version of the *Michigan Corpus of Upper Level Student Papers* (henceforth MICUSP_June09). More specifically, in line with the thematic focus of the present special issue, we chose to use the example of (un)attended *this* to exemplify how converging evidence from quantitative and qualitative corpus-linguistic methods and tools can provide a much more comprehensive picture of linguistic phenomena than either method could achieve alone. Recent work in corpus linguistics has emphasized the usefulness of quantitative and multifactorial analyses (cf. Gries 2003; Keune et al. 2005; Wulff 2008). While we fully subscribe to the value of a large-scale, quantitative perspective, we would like to demonstrate in this paper how a combination of results obtained from such a quantitative perspective can guide further qualitative analysis of the data, ultimately resulting in a more comprehensive, informative, and meaningful investigation of the phenomenon at hand.

More specifically, we used the case of (un)attended *this* to make this point by combining the findings of a Distinctive Collexeme Analysis and a logistic regression analysis, both representing a more quantitative perspective on the data, with an ensuing pattern analysis along more traditional lines of corpus-linguistic research. In a third step, we examined discipline, academic proficiency, and text-positional trends of some of the most prominent patterns identified. In line with the pattern approach adopted here, we focused on the sentential context in which tokens of *this* occurred. These local patterns, as we may call them, were expected to manifest themselves in different proportions of (un)attended *this* as far as academic disciplines are concerned, based on findings reported in Swales (2005). We were also interested in seeing whether there were significant variations in terms of writers' academic level, gender, or native speaker status.

The present paper is structured as follows. After a brief review of previous studies on (un)attended *this* in Section 2, Section 3 is devoted to an overview of the makeup of MICUSP_June09, and explains the different methodologies employed in the present study. The main results of the different analyses are presented in Section 4. Section 5 closes with a discussion of the main findings and their implications for EAP teaching and TESOL, as well as with some concluding remarks on avenues for future research.

2. *This* and its attendance: Some background issues

The issue of whether *this* should be attended by a following noun phrase has had a curiously muted academic and pedagogical history. Perhaps lurking in the deeper background is the well-known injunction in the ubiquitous Strunck and White (1979) to "omit needless words". Also in the background would appear to be the apparent belief among syntacticians and grammarians (e.g. Quirk et al. 1985; Biber et al. 1999; Huddleston and Pullum 2002) that a decision whether to follow a demonstrative with an NP (or not) is not a topic that falls within the purview of grammar, but rather one that belongs to stylistics, rhetoric or even information-processing.

One of the very few academic papers that directly focus on the topic is that by Geisler et al. (1985), which remains today the most sophisticated statement from a functionalist perspective. They stress the competing demands of economy and clarity: "Out of control, the unattended *this* points everywhere and nowhere; under control, it is the language's routine for creating a topic out of a central predication, pointing to it, bringing it into focus, and discussing it; all done in one stroke, gracefully, economically, and without names." (Geisler et al. 1985: 153)

Despite these strengths, the 1985 article, although published in a leading journal, seems to have been almost entirely neglected. As the reader may surmise, there is something of a mystery here. One of the very few papers to cite Geisler et al. is Finn (1995), who adopts an information-theoretical approach which argues that redundancy (adding unnecessary NPs following *this*) has an explicit cost: "Using more symbols to convey the same amount of information slows down the flow of new information to the reader." (1995: 244) Finn does suggest, however, that use of more interpretive NPs employing lexical items *not* present in the immediately previous text can have value. Similarly, Francis (1986) points out that an attending NP can be a powerful attitudinal signal from the author to the reader. After this, the scholarly trail goes cold, apart from a 2005 paper by Swales to be discussed later.

There is also surprisingly little coverage in all the textbooks and manuals designed to help U.S. students with their university writing tasks. Most, such

as Ede (2004), Faigley (2007) and Axelrod and Cooper (2008), merely make occasional comments about avoiding "vague reference" in connection with words like *this*, *it* and *which*, sometimes followed by illustrative revisions. As might be expected, there is somewhat more in Joseph Williams' well-regarded *Style: Ten Lessons in Clarity and Grace* (1985), in which he notes that sentence-initial nominalizations (containing a demonstrative) are an important way of realizing given-new patterns: "That is one important function of nominalizations: to sum up in one phrase actions you have just mentioned so that you can comment on them." (1985: 40) However, he does not discuss cases where unattended *this* might be warranted, or whether there are advantages in opting for more interpretive summary phrases. Technical communication textbooks also offer some general advice: "In almost all cases, demonstrative pronouns should be followed by nouns" (Markel 2004: 229); "Train yourself to avoid using 'It is . . .' and 'This is . . .' sentences. Occasionally, these sentences are fine, but some writers rely on them too much. You are better off minimizing their use in your writing." (Johnson-Sheehan 2005: A-10) Even so, neither author discusses cases where attending *this* with an NP may be unnecessary.

In the English for Academic Purposes field, the Swales and Feak textbooks (2000, 2004) give considerable attention to the lexical selection of nouns and noun phrases following a demonstrative in sentence-initial contexts, but they also do not discuss possible exceptions. Instead, they argue that non-native speakers of English should avoid unattended *this* both to reduce potential ambiguities and also to make a more professional impression on their readers. (Whether this is appropriate advice for non-native speakers of English is a matter we will return to in our closing comments.) After reviewing EAP work on the topic, Swales (2005) investigated the use of attended and unattended *this* in sentence-initial position in a subset of 80 research articles of the Hyland corpus, drawn from eight research fields (Hyland 1998). Given the foregoing discussion, the percentages of unattended *this* are considerably higher than occasional usage might suggest, ranging from a low of 25% in dentistry to a high of 56% in philosophy. He offers some preliminary explanations for these relatively high numbers for unattended *this,* one being that the absence of following noun phrases can be associated with main verbs that "are syntactically and semantically simple." (Swales 2005: 13)

One of the uncertainties about a published corpus of research articles is the influence of editors and reviewers on the style of the eventual finished text. We know, for example, that in very many cases, research article authors are required to shorten the length of the articles in order to fit them into journal or editorial requirements. Obviously, a word or two can be saved by omitting associated nominals following a demonstrative and, more importantly, can be saved without either elaborate rewriting or without upsetting the previously established given-new flow of information. With the MICUSP corpus, consist-

ing as it does of student papers with little need to conform to precise word limits, this particular uncertainty is avoided. Preliminary research on the data for *this* from this corpus (Römer and Wulff 2010) shows that (i) *this* is common and is in fact the eleventh most frequent word; (ii) the average percentage of attended *this* is higher (at 73%) than the research article average (64%) reported by Swales (2005); (iii) like in Swales, disciplinary variation in frequency use of unattended *this* – except for philosophy – is relatively muted; (iv) most intriguingly, attended *this* percentages increase slowly but consistently from the final year undergraduate to the third year graduate sub-corpora; and (v) the most common attending nouns are either metadiscoursal or related to methodology. However, Römer and Wulff did not address the key question of what might cause a quarter of the occurrences of *this* to be left unattended.

The traditional approach to this question, as pioneered by Geisler et al. (1985) and followed by all the manual writers, has been to explore possible ambiguities in the antecedent. This, at least in the case of Geisler et al., involved careful text analysis of a limited number of exemplars. However, with a large electronic corpus at hand, it is possible to explore the hypothesis that at least part of the answer lies not in the characteristics of the *preceding* text, but in the characteristics of the text that *follows* an occurrence of *this*. Geisler et al. note that the use or otherwise of an attendant noun would seem to be part of a writer's tacit understanding of how to write effective academic prose in English. With a corpus at hand, it may be possible to bring some of these tacit understandings into sharper focus – there may be patterns here that guide writers' choices.

3. Data and methods

3.1 *MICUSP: A new corpus of proficient student academic writing*

The Michigan Corpus of Upper-level Student Papers (MICUSP), compiled at the English Language Institute of the University of Michigan, Ann Arbor, is a new corpus of student academic writing samples.[1] The corpus, the first of its kind in North America, enables corpus researchers, EAP teachers, and testers to investigate the written discourse of proficient, advanced-level native and non-native speaker student writers at a large American research university. The corpus was made freely available to the global research and teaching community through an online search and browse interface in December 2009 (see http://search-micusp.elicorpora.info/).

MICUSP consists of 829 A-graded papers (totaling about 2.6 million words) of different types (e.g. research paper, report, response) from a wide range of different disciplines within four academic divisions, as listed in Table 1.[2] The

Table 1. *MICUSP composition: Distribution of papers across academic divisions and disciplines (figures based on June 2009 pre-release version)*

Academic division	Discipline	Papers	Tokens
Humanities & Arts	English	96	260,896
	History & Classical Studies	41	184,377
	Linguistics	38	108,117
	Philosophy	42	105,442
Social Sciences	Economics	25	65,883
	Education	47	143,432
	Political Science	62	199,507
	Psychology	103	315,200
	Sociology	68	192,521
Biological & Health Sciences	Biology	66	158,490
	Natural Resources & Environment	63	169,075
	Nursing	41	155,800
Physical Sciences	Civil & Environmental Engineering	29	72,898
	Industrial & Operations Engineering	42	101,255
	Mechanical Engineering	26	66,681
	Physics	21	38,695
Total		**810**	**2,338,269**

papers included in MICUSP were written by students of four different levels of study: final year undergraduates, and first, second, and third year graduate students. The corpus thus enables both analyses of disciplinary and developmental phenomena. Each of the papers in MICUSP has been marked up in TEI-compliant XML and maintains the structural divisions (sections, headings, paragraphs) of the original paper. A file header that has been added to each MICUSP file includes, among other things, information about the discipline and the student's level, native-speaker status, and gender, which makes it possible to carry out customized searches in subsections of the corpus, e.g. only in Biology papers written by native-speaker final year undergraduate students.

The analyses reported in this paper are based on a pre-release version of MICUSP compiled in June 2009 (MICUSP_June09). This version of the corpus consists of 810 student papers from 16 disciplines and four levels and contains around 2.3 million words (see Table 1 for a distribution of papers and tokens across disciplines). For the present study, the 810 files have been organized into subsets according to discipline and student level to enable targeted searches. Around 52.7% (427) of the papers included in MICUSP_June09 were written by final year undergraduates, and 47.3% (383) of the papers by first to third year graduate students (first year: 198 papers, 24.2%; second year: 112 papers, 13.8%; third year: 73 papers, 9%). Less than 20% (161) of the papers were produced by non-native speakers of English, while most papers

(649 altogether) of MICUSP_June09 have native-speaker authors. The ratio of female to male student authored papers is 500 to 310 (61.7% female; 38.3% male).[3]

For the purpose of the present study, we extracted all instances of *this* from MICUSP_June09 using *MonoConcPro 2.2* (Barlow 2004b). The resulting 15,711 hits were examined manually to identify instances in which *this* does not function as either a demonstrative determiner or pronoun. Five instances out of the 15,711 hits were uses of *this* as an intensifier (as in *The system is just this perverse*), leaving 15,706 hits in the sample. These instances were then coded for the position of *this* in the sentence. The 5,827 instances that were coded as sentence-initial cases of *this* constitute the data sample of the analyses to be described below.[4]

3.2 *Methods*

3.2.1 *Distinctive collexeme analysis*

In order to be able to address the question which verbs specifically are associated with (un)attended *this*, we computed a so-called *Distinctive Collexeme Analysis*. Distinctive Collexeme Analysis (DCA) is one member in the family of methods referred to as *collostructional analysis* (Stefanowitsch and Gries 2003; Gries and Stefanowitsch 2004), all of which measure the association between one linguistic construction (typically a verb) and another (in our example, attended and unattended *this*).[5] DCA is specifically tailored to identify the verbs that are significantly associated with attended *this* in direct contrast to unattended *this*, that is, distinguish best between attended and unattended instances. To test whether a given verb lemma does significantly distinguish between attended and unattended *this*, that is, qualifies as a distinctive collexeme of either variant, four frequencies are entered into a 2-by-2 table:

– the token frequency of that lemma with attended *this*;
– the token frequency of that lemma with unattended *this*;
– the overall frequency of attended *this*;
– the overall frequency of unattended *this*.

A Fisher-Yates exact text is applied to that table, providing a *p*-value which is, for ease of exposition, log-transformed to the base of ten and multiplied with −1.[6] A resulting *p*-value equal to or higher than approximately 1.3 corresponds to a probability of error of exactly or less than 5%; that is, it is statistically significant. The higher the log-transformed value, the higher the verb's distinctiveness. For both case studies, we first retrieved all relevant frequencies for all verb lemmas attested with attended and unattended *this* and then

computed the DCA with Coll.analysis 3.2 (Gries 2007). The results of the DCA are summarized in Section 4.2 below.

3.2.2 *Binary logistic regression*

We also computed a logistic regression analysis in order to identify the most influential determinants of the distribution of (un)attended *this* in our data, to see if and to what extent the choice of construction is indeed associated with the linkage between (un)attended *this* and the verb it occurs with as suggested by the DCA, and to identify possible interactions of determinants that would not surface in a monofactorial approach. Technically speaking, a binary logistic regression is used to determine the probability of an event that can surface in two distinct ways. In our case, we want to determine the probability of the predicted level of the dependent variable (which, for technical reasons, is set to be unattended *this*) on the basis of the following predictors:

1. the lemma frequency of the verb co-occurring with (un)attended *this* (**LOGFREQVERB**): lemma frequencies for all instances of (un)attended *this*. We used the British component of the *International Corpus of English* (ICE-GB) as a reference corpus to obtain the verb lemma frequencies. Next to the pragmatic advantage that the ICE-GB offers lemmatization, we wanted to obtain verb lemmas frequencies from a more balanced corpus than MICUSP represents (maybe most importantly including spoken language) as a better approximation of the cognitive entrenchment of these verbs (cf. Jurafsky 2003 for discussion of balanced corpora of even relatively small size like the Brown corpus correlate quite reliably with experimental data on word frequencies).

   For 71 verbs attested in the present data sample, no frequencies could be obtained from the ICE-GB, so unlike all other analyses to be presented below, the logistic regression is based on 5,756 instances of (un)attended *this*. Furthermore, for the logistic regression analysis, the lemma frequencies were logged.
2. the $p_{\log}$ values obtained from the DCA (described in detail in Section 3.2.2 below) indicating each instance's association strength with (un)attended *this* (**DISTINCTIVENESS**). For the logistic regression analysis, the $p_{\log}$ values were converted to a negative value if the verb is distinctively associated with unattended *this*.
3. the academic division (**DIVISION**) in which each instance occurred: this information was retrieved from the corresponding file header for each instance, resulting in predictor variable with four variable levels (henceforth abbreviated as 'biohealthsciences', 'humanitiesarts', 'physicalsciences', and 'socialsciences').
4. the academic proficiency level (**LEVEL**) of student writing in which each instance occurred, likewise retrieved from the corresponding file

headers; this predictor variable accordingly had four variable levels also (henceforth abbreviated as 'finalyearug', 'firstyeargrad', 'secyeargrad', and 'thirdyeargrad').

5. the gender (**GENDER**) of the student contributing each instances, also obtainable from the file headers, comprising two levels ('female' and 'male').
6. the native speaker status (**NATIVENESS**) of the student contributing each instance, again retrieved from the corresponding file headers, and also comprising two levels ('native' and 'nonnative').

A logistic regression works as follows: in a first step, all predictors (that is, variables and their potential two-way interactions) are entered into a logistic model. On the basis of model comparisons (using the function ANOVA in *R*), the predictor (starting from the highest level of interactions) that makes the least significant contribution to the model is identified and discarded, and another logistic model is computed without this predictor. This model fitting process is performed iteratively until only significant predictors remain in a final model, which is also referred to as a minimal adequate model.[7,8]

3.2.3 *Identification of common* this-*clusters*

In order to complement the quantitative DCA and logistic regression analysis, we carried out a more qualitatively oriented analysis which focuses on phraseological items in MICUSP. This analysis takes a closer look at recurring multiword units with the word *this* (e.g. *This means that*, *This is not to say*), their distribution, and their functions in advanced student writing across disciplines.

We started our phraseological analysis by extracting from MICUSP_June09 *this*-clusters of different spans, i.e. contiguous word sequences that contain the word *this*. The tool we used for this cluster extraction is *Collocate* (Barlow 2004a), a software package that retrieves lists of n-grams of different lengths and of collocations (or clusters) with a specific search word in a set span from a text or corpus. To create lists of *this*-clusters, we used the *Collocate* "Word/Phrase Extract" function. We carried out both case-sensitive and case-insensitive searches for spans of two to six words (e.g. *this paper*, *this means that*, *this seems to be*, *this is due to the*, *this is not to say that*). We discuss the resulting frequency-sorted lists of *this*-clusters in Section 4.3 below.

3.2.4 *Analysis of the distribution of selected* this-*clusters across disciplines, levels and texts*

In a next analytic step, we examined concordances of selected high-frequency *this*-clusters from the *Collocate* cluster lists, focusing on *this* + verb clusters (e.g. *this means*, *this implies that*) identified previously by means of the DCA. For these prominent *this* + verb clusters, we checked how they are distributed

(a) across disciplines, (b) across student levels, and (c) across texts. A MICUSP n-gram database designed by Matthew O'Donnell (O'Donnell and Römer in preparation) enabled us to identify how often each cluster occurs, in which of the 16 MICUSP disciplines and four levels, and whether it prefers (or avoids) certain positions in the sentence, paragraph, or text.[9] While relations between language items and text structure – in Hoey's (2005) terms *textual colligations* – have been extensively studied in the language of newspapers (see e.g. Hoey 2005, 2009; Hoey and O'Donnell 2008; Mahlberg and O'Donnell 2008), they are now being examined in spoken and written academic discourse as well (Csomay 2009; Römer 2010, respectively). Hence, this part of the analysis rounds off our attempts to combine fairly novel corpus-linguistic techniques.

4. Results

Out of 5,827 instances in the data sample, 2,499 (43%) are cases of unattended and 3,328 (57%) are cases of attended *this*. Considering the above-mentioned stylistic cautions against unattended *this* in academic writing, unattended *this* therefore occurs much more frequently than may have been expected. (Discipline-, proficiency level-, gender-, and nativeness-specific distributions are provided in Tables 6–9 in the appendix.)

4.1 *Distinctive collexeme analysis*

Tables 2 and 3 provide a summary of the results of the DCA. Looking at the collexemes distinctive for unattended *this* (Table 2), we find that by far the most distinctive verb lemma is *be*, followed by other high frequency, semantically bleached verbs *mean* and *do*. This result gains even more significance once we take into consideration that according to the logistic regression results, high verb lemma frequency generally pulls towards attended *this*. As to the verbs further down in the ranking, one property many of them share is that they are mostly used to signal upcoming commentary on or discussion of some previously described process or result (*lead*, *result*, *happen*, and *attribute* are examples in question).

When we turn to the collexemes distinctive for attended *this* (Table 3), a much more diverse picture emerges. While the list of distinctive collexemes for attended *this* is much more extensive than that for unattended *this*, we also see that none of the verb lemmas distinctive for attended *this* reaches a *p*-value in the same range as *be* and *mean* do for unattended *this*. In terms of a general semantic trend in these verbs, we can make out a comparatively more pronounced preference for verbs that are typically used to initiate the description

Table 2. *Collexemes distinctively associated with unattended sentence-initial* this *in MICUSP_June09*

Verb lemma	FYE_{log}
be	103.631
mean	23.027
do	6.055
lead	5.928
result	3.495
happen	2.412
attribute	2.289
leave	1.944
imply	1.835
seem	1.694
accomplish	1.599
fall	1.562
measure	1.472
allow	1.376
increase	1.358
cause	1.355

or the structural outline of a paper or study (*examine*, *focus*, *explore*, and *investigate*), or to refer to methodology (*use*).

4.2 *Multi-factorial analysis: logistic regression*

The minimal adequate model shows that there is a highly significant strong correlation between the predictors listed in Table 4 below and the choice of (un) attended *this* (log-likelihood ratio $\chi^2 = 1282.08$; df = 22; $p = 0$). Nagelkerke's R^2, an indicator of general correlational strength between the predictor and the dependent variable, amounts to .268, and the model has near-good classificatory power: C = .771; Dxy = .541 (usually, a C value of 0.8 or higher is considered "good"; see Harrell 2001) . On the basis of the minimal adequate model, 69.82% of all instances can be correctly predicted as either attended or unattended (the random classification accuracy amounts to 56.80%).

Table 4 lists the significant predictors of the minimal adequate model in descending order of their (absolute) coefficient values as obtained from the lrm output (see endnote 8; while Table 4 only contains (marginally) significant predictor levels, a more complete overview is given in Table 10 in the appendix, which lists all predictor levels, standard errors, Wald's *z*, and confidence intervals). The reference level of the independent variable is unattended *this*, so positive coefficients indicate a positive correlation with unattended *this* (or conversely a negative correlation with attended *this*), while negative coefficients

Table 3. *Collexemes distinctively associated with attended sentence-initial* this *in MICUSP_June09*

Verb	FYE_{log}	Verb	FYE_{log}
use	8.697	reveal	2.133
examine	8.275	serve	2.028
focus	7.321	highlight	1.988
find	6.586	take	1.959
explore	4.631	define	1.947
base	4.386	know	1.947
seek	3.898	propose	1.947
provide	3.446	rely	1.947
contain	3.410	begin	1.921
investigate	3.410	become	1.817
have	3.390	cover	1.704
discuss	3.218	design	1.704
aim	3.166	consider	1.536
consist	3.166	associate	1.460
perform	3.166	exist	1.460
review	2.922	look	1.460
attempt	2.782	remove	1.460
describe	2.782	treat	1.460
show	2.722	place	1.425
compare	2.435	apply	1.392
intend	2.435	illustrate	1.380
support	2.411	address	1.300
present	2.199	analyze	1.300
receive	2.191	argue	1.300
continue	2.182	draw	1.300

Table 4. *Significant predictors of the minimal adequate logistic regression model*

Predictor	Coeff.	*p*
DISTINCTIVENESS	−1.032	0.000
DISCIPLINE = physicalsciences × LEVEL = thirdyeargrad	−1.015	0.001
DISCIPLINE = humanitiesart × LEVEL = thirdyeargrad	−0.589	0.025
DISCIPLINE = biohealthsciences × GENDER = male	−0.331	0.064
LEVEL = firstyeargrad	−0.323	0.011
LOGFREQVERB × DISTINCTIVENESS	0.186	0.000
GENDER = male	0.202	0.054
LOGFREQVERB	0.225	0.000
DISCIPLINE = biohealthsciences × LEVEL = firstyeargrad	0.471	0.017
DISCIPLINE = humanitiesarts × LEVEL = firstyeargrad	0.827	0.000

indicate a negative correlation with unattended *this* (or conversely, a positive correlation with attended *this*).

As we can see in Table 4, by far the most significant predictor is DISTINCTIVENESS: the higher the DISTINCTIVENESS score, the less likely unattended *this* becomes (coefficient: −1.032). That is, we see that the results of the DCA are confirmed: verbs highly distinctive for unattended *this* (which were converted into negative p_{log} values for the purpose of the logistic regression) are very highly correlated with unattended *this*, and verbs highly distinctive for attended *this* (the values for which remained positive in the logistic regression) are very highly correlated with attended *this*. The fact that DISTINCTIVENESS yields the highest coefficient value lends strong support to our hypothesis that (un)attended *this* forms local patterns with its immediate lexico-syntactic environment.

Secondly, we see that DISCIPLINE and LEVEL interact in quite intricate ways. Both in the Humanities and Biological Health Sciences, we see at the bottom of Table 4 that first year graduate students significantly prefer to leave *this* unattended (coefficients: 0.827 and 0.471, respectively), which stands in contrast to the general tendency of first year graduate students to prefer attended *this*, as indicated by the significance of this predictor level (coefficient: −0.323). Third year graduate students, particularly in the Humanities and the Physical sciences, use attended *this* even more often (coefficients: −1.015 and −0.589, respectively). In sum, it seems that while there is a trend towards attended *this* already in first year graduate writing, this preference is even more pronounced in the third year, especially in the Physical Sciences, and students in the Humanities undergo the most dramatic development, starting out with a preference for unattended *this*, and ending up with a clear preference for attended *this*. Figure 1 provides a graphical display of this interaction.

In the Biological Health Sciences, Table 4 furthermore reveals an interaction with GENDER: male students in this discipline use attended *this* significantly more often than female students (coefficient: −0.331). Figure 2 is a graphical representation of this interaction. This strong preference for attended *this* by male students in the Biological and Health Sciences gains even more weight when seen in contrast to the general preference of male students across all disciplines to use unattended *this* significantly more often than their female classmates (coefficient: 0.202).

Finally, Table 4 shows that LOGFREQVERB is also a significant predictor: the more frequent the verb, the more likely unattended *this* becomes. The leftmost bar plot in Figure 3 displays this general trend graphically: for verbs with a logged frequency of 5 or higher, there is a clear incline in occurrence with unattended *this*. While LOGFREQVERB yields a significant result, we also see, however, that its effect can be overridden by the verb's DISTINCTIVENESS, as evidenced by the significant interaction between the two predictors: some

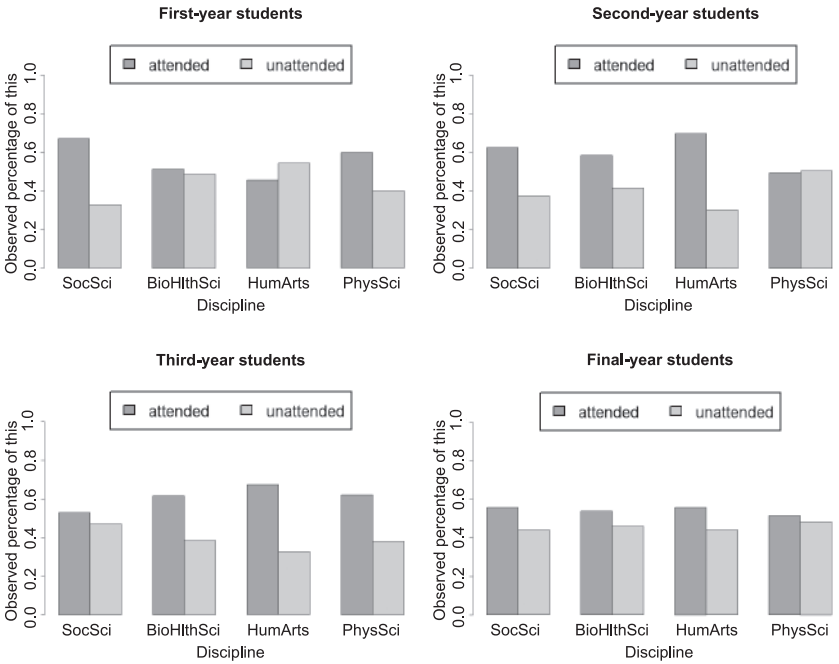

Figure 1. *Interaction between DISCIPLINE and LEVEL*

verbs are frequent but distinctively associated with attended *this*. Accordingly, the right bar plot in Figure 3 is a visual match to the middle bar plot which displays the main effect of DISTINCTIVENESS, while standing in a chiastic relationship to the left bar plot that displays the main effect of LOGFREQVERB.

4.3 *Common this-clusters*

Table 5 presents an overview of the 20 most frequent sentence-initial *this*-clusters of spans two, three, and four together with their frequencies of occurrence in MICUSP_June09.

As we can see, the 20 two-word *this*-clusters can be divided into three groups: *this* + modal verb, *this* + non-modal verb, and *this* + noun. Particularly frequent in the first category are the clusters *this would*, *this can*, and *this will*, while forms of *be* (*is*, *was*) top the overall list of items that immediately follow sentence-initial *this*. Among the nouns that most commonly form two-word clusters with *this* in our student writing corpus are *paper*, *study*, *model*, and *process*, so students often refer to either their own or other scholars' written or

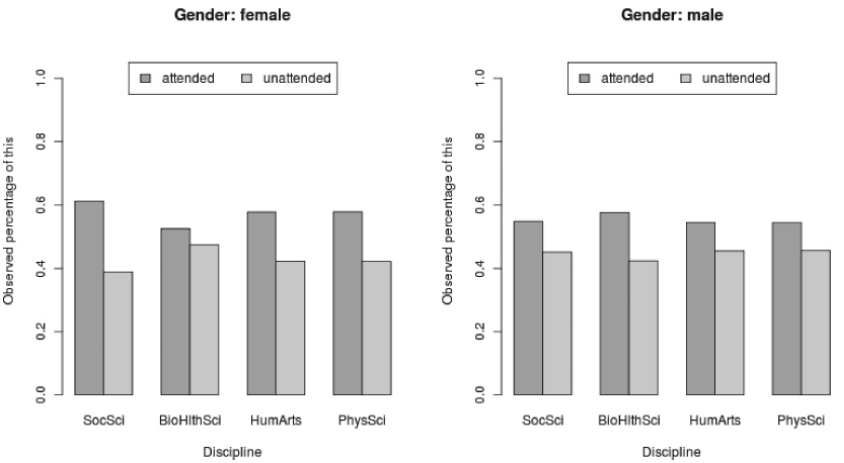

Figure 2. *Interaction between DISCIPLINE and GENDER*

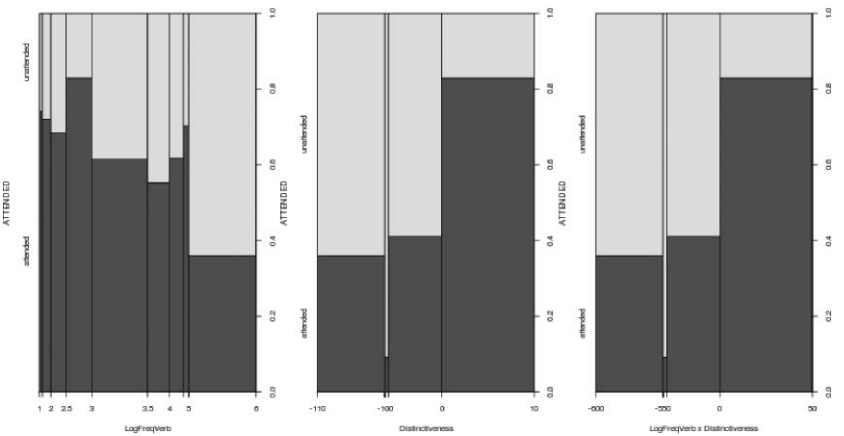

Figure 3. *Interaction between LOGFREQVERB and DISTINCTIVENESS*

empirical work. As for three-word clusters with *this*, we observe highest numbers for items that are extensions of common two-word clusters, such as *this is a*, *this is the*, *this can be*, and *this paper will*. Interesting also are two general patterns among these top-20 clusters: *this* + modal verb + *be* (e.g. *this can be*, *this could be*) and *this* + present tense form of a lexical verb + *that* (e.g. *this means that*, *this suggests that*) – both connected to the discourse function of

Table 5. *Top 20 most frequent* this*-clusters of different spans in MICUSP_June09 (Collocate output)*

Span 2	n	Span 3	n	Span 4	n
This is	711	This is the	64	This is not to	22
This paper	119	This is a	58	This is because the	18
This was	106	This is not	52	This means that the	16
This would	91	This can be	51	This is due to	13
This can	89	This is because	49	This can be seen	12
This will	79	This means that	45	This seems to be	11
This study	74	This paper will	41	This leads to a	10
This could	70	This is an	40	This is an important	10
This may	69	This could be	29	This suggests that the	9
This means	55	This may be	26	This could be due	8
This has	37	This type of	25	This implies that the	8
This suggests	34	This suggests that	23	This is important because	8
This process	33	This leads to	21	This can be done	7
This method	29	This implies that	20	This is a very	6
This model	28	This is in	20	This is especially true	6
This seems	28	This is important	19	This is one of	6
This leads	26	This seems to	16	This is similar to	6
This type	26	This would be	16	This leads to the	6
This argument	24	This was done	14	This is an interesting	4
This section	19	This will be	14	This focus on the	3

explaining. Most common among the span four *this*-clusters, too, are items that serve to provide or introduce explanations: *this is not to*, *this is because the*, *this is due to*.[10]

In the following section, we will focus on *this* + verb clusters and their patterns. These clusters turned out to be particularly frequent among all sentence-initial *this*-clusters in MICUSP_June09 (see Table 5).

4.4 *Distribution of selected* this *+ verb clusters across disciplines, levels, and texts*

For the analysis of disciplinary, level, and positional variation, we selected the following six *this* + verb clusters: *this is*, *this means*, *this leads*, *this implies*, *this seems*, and *this allows*.[11] All six combinations exhibit a high degree of morphological fixedness: attestations are predominantly in third person singular simple present tense, and at least in the sentence-initial position examined here, the verbs in these six clusters are predominantly unattended (and rank correspondingly high in the DCA). Concordance analyses of the six selected verb forms (with *this* occurring in up to five positions to the left) showed that

these verb forms are used in unattended contexts in 63.5% to 98.4% of all examined cases. A case in point here is the form *means*, for which only one attended example (1.6%) could be identified in MICUSP_June09 (*This size-selectiveness means . . .*). For the other five verbs, the shares of attendedness range from 18.1% (*leads*) to 36.5% (*is*). Among the nouns that follow sentence-initial *this* in these cases are general high-frequency academic nouns (e.g. *model*, *paper*, *method*, *process*, and *finding*) and technical terms such as *channel*, *equation*, *disparity*, and *varying treatment*. In the following, we will focus on unattended instances of sentence-initial *this* plus *is/means/leads/implies/seems/allows*, which account for the majority of the cases.

Starting with the by far most frequent *this* + verb cluster, *this is*, we observe some very interesting distributional trends, especially in terms of disciplinary and positional variation. While in most disciplines sentence-initial *this is* occurs between 28 and 38 times per 100,000 words of text, the normalized numbers of occurrence in Philosophy (71.04) and Physics (109.21) are much higher. A concordance analysis shows that in Philosophy papers, *this is* frequently appears in phrases that help to express explanations or give reasons for something, such as *this is (mainly) because*, and *this is why/how/what*. In Physics papers, patterns like *this is due to*, *this is (clearly) true*, and *this is an interesting result/a simple equation*, are used to make factual observations or explain findings. Generally common across all MICUSP disciplines are the patterns *this is because*, *this is due to*, *this is not*, *this is important (because)*, and *this is an important/interesting X*. *This is*-clusters are frequent across all four student levels, with highest figures found for the senior undergraduate and first-year graduate level and lower frequencies observed for the second- and third-year graduate level. As becomes apparent from the graph in Figure 4, this cluster shows a dispreference for text- and in particular paragraph-initial positions and occurs relatively much more often in the middle and final sections of paragraphs and texts (interestingly, attended instances, e.g. *this process/model/paper is*, show a roughly even distribution across paragraphs and texts, which may indicate that these instances perform different textual functions from the cluster *this is*).

A similar picture emerges in terms of positional variation when we look at sentence-initial *this means*. Again, paragraph- and text-initial positions are avoided while medial and final positions are preferred (Figure 5). Advanced student writers use this cluster most often in the middle of a paragraph to explain or rephrase something they stated at the beginning. Another common function of *this means*, or *this means that*, is to make predictions, as in the following example taken from a Mechanical Engineering paper: *This means that a higher frequency will be able to pump more heat into a room* [. . .]. *This means* is very unevenly distributed across MICUSP disciplines. With only 0.8 instances per 100,000 words, it is very rare in Biology and English,

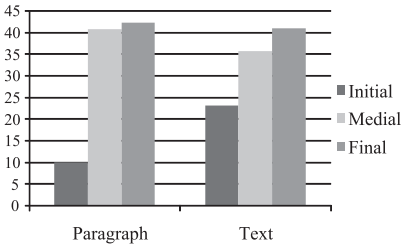

Figure 4. *Distribution of sentence-initial* this is *across paragraphs and texts in MICUSP_June09 (figures normalized per 100,000 words)*

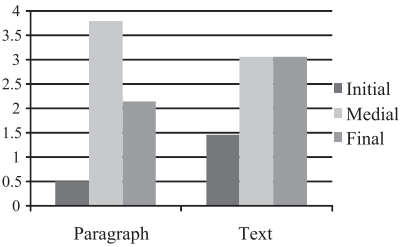

Figure 5. *Distribution of sentence-initial* this means *across paragraphs and texts in MICUSP_June09 (figures normalized per 100,000 words)*

whereas it is quite common in Mechanical Engineering (9.22) and Industrial & Operations Engineering (11.13). The cluster is about twice as common in senior undergraduate and second-year graduate writing as in first- and third-year graduate student writing, which can in part be explained by the fact that there are larger numbers of Engineering papers in the former two datasets.

An effect of discipline can also be observed in the distribution of sentence-initial *this leads*, usually followed directly by *to* and a noun phrase or verb infinitive, in some cases by a personal pronoun or name and then *to* and a noun phrase or verb infinitive (e.g. *This leads us to ask the following questions*). *This leads* is comparatively frequent in Physics and Nursing papers (6.07 and 4.49 per 100,000 words), rare in English and Education papers (0.82; 0.8), and not used at all in Civil & Environmental Engineering, History & Classical Studies, Industrial & Operations Engineering, and Natural Resources & Environment, where either no (predominantly negative) consequences are described, or they are described in different ways. There is hardly any variation across levels for this cluster, apart from a slightly lower result for first-year graduate level than for the three other levels. With respect to positional variation, we find that sentence-initial *this leads* (similar to *this is* and *this means*) avoids text-initial and (though to a lesser extent) paragraph-initial positions, while favoring text-medial and text- and paragraph-final positions (see Figure 6).

Sentence-initial *this implies*, generally followed by *that*, also shows clear preferences in terms of textual and disciplinary distribution, and occurs more often in third-year graduate student papers than in papers written by students on the three lower levels, which may mean that this cluster is associated with more advanced academic student writing. *This implies* is considerably more frequent in Physics and Economics papers (6.07 and 5.58 hits per 100,000 words) than in, for instance, Nursing or Philosophy papers (0.75; 1.03). The

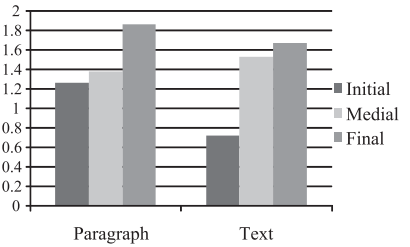

Figure 6. *Distribution of sentence-initial* this leads *across paragraphs and texts in MICUSP_June09 (figures normalized per 100,000 words)*

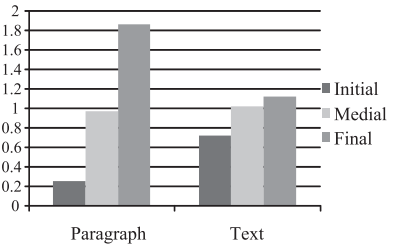

Figure 7. *Distribution of sentence-initial* this implies *across paragraphs and texts in MICUSP_June09 (figures normalized per 100,000 words)*

cluster does not at all occur in the MICUSP_June09 Education, English, Linguistics, Mechanical Engineering, and Sociology subsections. As a look at a concordance of the cluster indicates, *this implies (that)* functions to introduce important summarizing aspects or consequences of what has been previously discussed in the text. A typical example from a Natural Resources & Environment paper is: *This implies that the organism will do well in variable habitats*. As Figure 7 shows, *this implies* is clearly a paragraph-final cluster, which reflects its summarizing function. It also tends to occur more often in text-final and text-medial position than at the beginning of texts.

Another *this* + verb cluster that is very unevenly distributed across MICUSP disciplines is *this seems*. For this cluster, which mainly functions as a hedging or softening device, we find the highest number of hits by far (11.32 per 100,000 words) in Philosophy and rather low frequencies in Nursing (0.75) and Industrial & Operations Engineering (1.15). Interesting also in this context is the very high number of sentence-initial *it seems* in the MICUSP_June09 Philosophy subsection. There are 41.18 hits per 100,000 words, which accounts for around 45% of all instances of sentence-initial *it seems* in the corpus. While Philosophy shows an overall above-average share of unattended *this* (57%, see Table 6 in the appendix), the high frequencies found for *seems*-clusters cannot be solely explained on the basis of disciplinary preferences for unattended *this* patterns. Sentence-initial *this seems* does not occur in any of the papers from Biology, Education, History and Classical Studies, Mechanical Engineering, Physics, Political Sciences, and Sociology – perhaps because most of these disciplines are concerned with observable facts and discrete events. Students on all four levels use this cluster, with highest shares observed for second-year graduate student papers. Common patterns found with sentence-initial *this seems* include *this seems to (be)* and *this seems like*, the latter of which sounds somewhat colloquial and occurs exclusively in senior undergraduate and first-

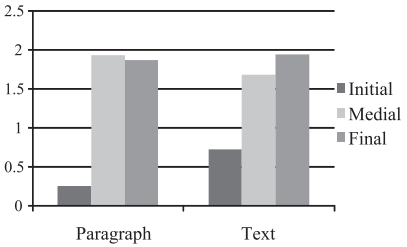

Figure 8. *Distribution of sentence-initial* this seems *across paragraphs and texts in MICUSP_June09 (figures normalized per 100,000 words)*

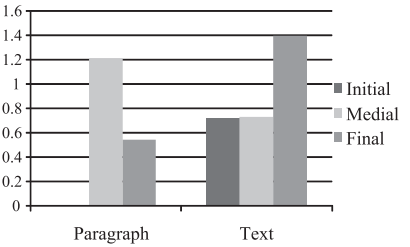

Figure 9. *Distribution of sentence-initial* this allows *across paragraphs and texts in MICUSP_June09 (figures normalized per 100,000 words)*

year graduate student papers. (In fact, this particular pattern is unattested in the Hyland corpus of published research articles.) *This seems* is, as Figure 8 illustrates, yet another *this* + verb cluster that avoids paragraph- and text-initial positions and prefers to occur somewhere in the middle or towards the end of a text or paragraph.

Finally, the 20 instances of sentence-initial *this allows* in MICUSP_June09 are fairly evenly distributed across ten out of the 16 MICUSP disciplines (with no hits in Civil & Environmental Engineering, Economics, History & Classical Studies, Linguistics, Physics, and Sociology). As to cross-level variation, the figures go down slightly with increasing writing proficiency, but absolute numbers are too small to justify any related conclusions. The cluster never occurs in the first sentence of a paragraph, is mainly paragraph-medial, and predominantly text-final (see Figure 9).

In sum, the semantic differentiation hinted at in the DCA is confirmed by the distributional analysis: The particular clusters analyzed in more detail above, which contain verbs distinctively associated with unattended *this*, can be further associated with their use as textual markers of upcoming interpretation, evaluation, and discussion; this is reflected in their positional preferences at the end of paragraphs and texts. Correspondingly, *this* + noun + verb clusters predominantly occur in text-initial and -medial positions, which again is in line with the semantic tendency for these clusters to initiate structural outlines and procedural descriptions. Figure 10 illustrates this point for *this* + noun + verb clusters that contain any of the verbs (in 3rd person singular form) listed in Table 3 as the most highly distinctive collexemes for attended *this*. While space does not permit a detailed presentation of all results, suffice it here to say that for the cluster containing verbs highly distinctively associated with attended *this*, we can observe highly similar positional preferences throughout.

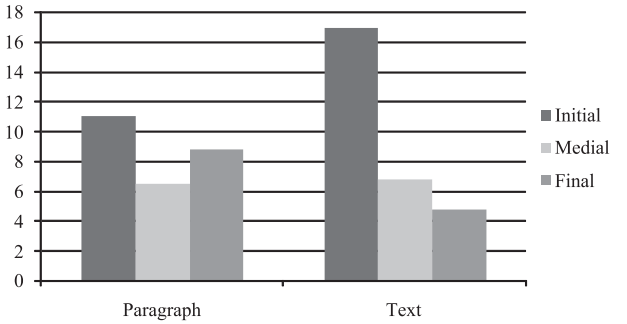

Figure 10. *Distribution of sentence-initial* this * + *3rd person singular forms of verbs listed in Table 3* (e.g., this * examines, this * focuses, this * explores) *across paragraphs and texts in MICUSP_June09 (figures normalized per 100,000 words)*

5. Discussion and conclusion

By combining a quantitative and a qualitative perspective on (un)attended *this* in academic student writing, various hitherto unexamined properties of this alternation have been identified. Firstly, while the binary logistic regression turned out to have moderate to good predictive power overall, it strongly suggests that the choice between attended and unattended *this* hinges in part on the choice of verb that accompanies it. As the DCA showed, this lexical drive is particularly pronounced for the most frequent cases of unattended *this* (*this is* and *this means*), a result that stands out, especially given the general, statistically significant positive correlation between high frequency verb lemmas and attended *this*. The binary logistic regression furthermore revealed that unattended *this* is somewhat preferred by male authors (with the exception of male students in the Biological Health Sciences – a finding that calls for further research). Thirdly, the logistic regression analysis confirmed a significant bias towards attended *this* already in first year graduate student writing, but students in the Humanities only develop this preference over time (and ultimately exhibit an even stronger preference for attended *this* than their fellow students except for the Physical Sciences).

Beyond the semantically bleached present tense sentence-initial uses of *be* and *mean*, the DCA helped us to uncover a semantic bias of verbs co-occurring with unattended *this* as mainly oriented towards commentary, evaluation, or discussion. While the DCA attests a partial overlap of unattended and attended *this* with regard to this class of verbs (such as *discuss*, *argue*, and *review*), the DCA furthermore revealed that a considerable share of the verbs distinctively associated with attended *this* are largely descriptive and oriented toward indicating structural outlines, the latter apparently being associated with attended

this only. As our pattern analyses confirmed, these lexical biases resurface at the text-organizational level, as evidenced in corresponding paragraph- and text-positional preferences.

In combination, the results obtained from these different points of view all suggest that certain sentence-intial *this* + verb clusters form relatively fixed contiguous patterns that can be considered new units of meaning in which the two components (*this* and the verb form) take on a meaning that is different from the meaning they take on in other contexts. More specifically, it appears that the six *this* + verb clusters we looked at more closely (with *this is* clearly taking the lead) function as interpretative markers signaling the transition from observation and description to summary, interpretation, and evaluation both within a given paragraph and from a textual perspective. This finding resonates with Sinclair's (2004) observations on the *Idiom Principle* and the "phraseological tendency" of language, i.e., the fact that words do not appear in isolation but "go together and make meanings by their combinations." (p. 29)

These findings have useful implications for those who teach courses in or write textbooks on academic writing. As the discussion in Section 2 has intimated, the general consensus both among those who teach native speakers and those who teach or write material for non-native speakers of English would be to advocate against employment of unattended *this*. For example, Swales and Feak (2000, 2004) argue that following *this* with an appropriate NP (i) reduces potential comprehension problems on the part of the reader and (ii) can make the writer appear more professional and authoritative. Whatever the merits of these arguments, the data presented in this paper suggests very strongly that there are high frequency phrases such as *This is because* and *This means that* which need to be noted as valid exceptions to any general advice and then incorporated into teaching materials for apprentice academic writers.

In Section 2, we raised the issue as to whether the topic of this paper falls within the purview of grammar, stylistics, rhetoric, or information processing. With the analysis completed, we now briefly return to this topic. As a matter of practice, the sparse literature on (un)attended *this* underscores its orphan status, unrecognized as legitimate by any of its potentially parent disciplines. Since it rarely, if ever, surfaces in either of the large bodies of work devoted to English grammar and to the information processing of English prose, we are left with stylistics or rhetoric. While it is true that unattended *this* is somewhat more frequent in spoken contexts (as in the MICASE research sub-corpus; cf. Swales 2004), it also remains a common occurrence in our MICUSP data (43% of the total instances of sentence-initial *this*). As there are no decisive correlations with either level of formality or with writing as opposed to speech, we are left with rhetoric. This conclusion makes some sense in that – apart from the formulaic patterns we have uncovered – writers doubtless juggle Geisler et al.'s (1985) competing claims of clarity versus economy. In the end, the topic finds

at least an occasional home in that part of rhetoric that concerns itself with audience analysis.

On a final note, the overall predictive power of the logistic regression analysis cautions us that while the strong verb-specific associations with (un)attended *this* are solidly confirmed by our analyses, in terms of cause and effect, the verb need not necessarily be the first link in the chain driving the choice between attended and unattended *this*. Since the present study did not take the antecedents of *this* into consideration, it cannot yet be ruled out that the choice of verb is indeed a consequence of the writer's choice of (un)attended *this* that ultimately depends more crucially on the nature of the noun phrase referent. In other words, it is conceivable that the choice for either attended or unattended *this* is made before the verb is selected.

Ultimately, addressing this issue requires more extensive analyses of the data at various levels (many of which become obvious only after doing the kind of quantitative and qualitative analyses we have presented here). We plan to expand our current analysis not only with regard to the characteristics of the noun phrase and their antecedents, but also in terms of the positional variation of (un)attended *this* in a given sentence; the morphological characteristics of the verb in terms of tense, aspect, person, and number marking; and differences between *this* and its competitors *that*, *these*, and *those*. Moreover, it would be interesting to compare the results gleaned from academic student writing with expert academic writing. Finally, the exploration of the text-distributional characteristics of *this*-clusters presented in Section 4.4, which was motivated in large parts by the findings of the preceding quantitative analyses, could in turn be followed up by another more quantitatively-minded analysis, for instance in the form of a Poisson regression with the observed cluster frequencies as the dependent and textual position and verb as the independent variables.

Appendix

Table 6. *Distribution of (un)attended sentence-initial* this *across disciplines in MICUSP_June09*

Academic division	Discipline	Unattended *this*	%	Attended *this*	%
Humanities & Arts	English	174	35	329	65
	History & Classical Studies	134	37	229	63
	Linguistics	144	51	138	49
	Philosophy	175	57	133	43
Social Sciences	Economics	90	43	118	57
	Education	135	41	195	59
	Political Science	176	37	296	63
	Psychology	342	42	480	58
	Sociology	182	40	271	60

Table 6. (*Continued*)

Academic division	Discipline	Unattended *this*	%	Attended *this*	%
Biological & Health Sciences	Biology	129	37	224	63
	Natural Resources	207	51	200	49
	Nursing	181	49	190	51
Physical Sciences	Civil & Environmental Engineering	83	32	178	68
	Industrial & Operations Engineering	160	47	181	53
	Mechanical Engineering	105	52	98	48
	Physics	82	55	68	45
Total		2,499	43	3,328	57

Table 7. *Distribution of (un)attended sentence-initial* this *across levels in MICUSP_June09*

Proficiency level	Unattended *this*	%	Attended *this*	%
Final year undergraduate	1,144	45	1,416	55
First year graduate	727	43	960	57
Second year graduate	361	39	575	61
Third year graduate	267	41	377	59
Total	2,499	43	3,328	57

Table 8. *Distribution of (un)attended sentence-initial* this *by gender in MICUSP_June09*

Gender	Unattended *this*	%	Attended *this*	%
Female	1,396	42	1,954	58
Male	1,103	45	1,374	55
Total	2,499	43	3,328	57

Table 9. *Distribution of (un)attended sentence-initial* this *by native speaker status in MICUSP_June09*

Native speaker status	Unattended *this*	%	Attended *this*	%
Non-native speaker	499	45	607	55
Native speaker	2,000	42	2,721	58
Total	2,499	43	3,328	57

Table 10. *Complete output of the minimal adequate logistic regression model*

Predictor	Coeff.	S.E.	Wald's *z*	*p*	2.5% CI	97.5% CI
DISTINCTIVENESS	−1.032	0.055	−18.730	0.000	−1.142	−0.926
DISCIPLINE = physicalsciences × LEVEL = thirdyeargrad	−1.015	0.312	−3.250	0.001	−1.631	−0.407
DISCIPLINE = humanitiesart × LEVEL = thirdyeargrad	−0.589	0.263	−2.240	0.025	−1.108	−0.078
DISCIPLINE = biohealthsciences × GENDER = male	−0.331	0.179	−1.860	0.064	−0.682	0.018
LEVEL = firstyeargrad	−0.323	0.126	−2.550	0.011	−0.572	−0.076
DISCIPLINE = biohealthsciences × LEVEL = thirdyeargrad	−0.285	0.317	−0.900	0.370	−0.912	0.334
DISCIPLINE = humanitiesarts × GENDER = male	−0.221	0.165	−1.340	0.182	−0.544	0.103
LEVEL = secyeargrad	−0.215	0.141	−1.530	0.126	−0.492	0.059
DISCIPLINE = humanitiesarts × LEVEL = secyeargrad	−0.194	0.232	−0.840	0.402	−0.651	0.259
DISCIPLINE = humanitiesarts	−0.056	0.125	−0.450	0.653	−0.301	0.189
DISCIPLINE = physicalsciences × LEVEL = firstyeargrad	−0.011	0.220	−0.050	0.961	−0.443	0.421
DISCIPLINE = physicalsciences	0.042	0.180	0.230	0.815	−0.312	0.393
DISCIPLINE = physicalsciences × LEVEL = secyeargrad	0.105	0.254	0.420	0.678	−0.392	0.603
DISCIPLINE = biohealthsciences × LEVEL = secyeargrad	0.160	0.250	0.640	0.522	−0.332	0.650
LOGFREQVERB × DISTINCTIVENESS	0.186	0.010	18.580	0.000	0.167	0.206
DISCIPLINE = biohealthsciences	0.192	0.134	1.430	0.154	−0.072	0.455
GENDER = male	0.202	0.105	1.930	0.054	−0.003	0.408
LEVEL = thirdyeargrad	0.213	0.143	1.490	0.137	−0.068	0.495
LOGFREQVERB	0.225	0.050	4.490	0.000	0.127	0.324
DISCIPLINE = physicalsciences × GENDER = male	0.275	0.202	1.360	0.174	−0.121	0.673
DISCIPLINE = biohealthsciences × LEVEL = firstyeargrad	0.471	0.198	2.380	0.017	0.083	0.860
DISCIPLINE = humanitiesarts × LEVEL = firstyeargrad	0.827	0.195	4.250	0.000	0.447	1.210

Bionotes

Stefanie Wulff is an Assistant Professor in the Department of Linguistics and Technical Communication at the University of North Texas. She received her Ph.D. from the University of Bremen, Germany, in 2007, and has held a post-doctoral fellowship at the University of Michigan and a lecturer position at the University of California at Santa Barbara. Her research interests are in the

areas of quantitative corpus linguistics, construction grammar, second language acquisition, and student writing. Email: Stefanie.Wulff@unt.edu

Ute Römer is currently an Assistant Professor at Georgia State University (GSU). Prior to joining GSU in 2011, she was the director of the applied corpus linguistics unit at the University of Michigan English Language Institute where she managed the Michigan Corpus of Upper-level Student Papers (MICUSP), among other projects. Her primary research interests include corpus linguistics, phraseology, academic discourse analysis, and the application of corpora in language learning and teaching. Her current research focuses on student academic writing, on how corpus tools and methods can be used to identify meaningful units in specialized discourses, and on combining corpus- and psycholinguistic evidence to gain insights into speakers' use and acquisition of English verb-argument constructions. Email: uroemer@gsu.edu

John Swales is professor Emeritus of Linguistics at the University of Michigan, where he was also Director of the English Language Institute from 1985 to 2001. Publications in 2011 include *Navigating academia: Writing supporting genres* (with Christine Feak) and a reissue of the 1981 monograph Aspects of article introductions, both published by the University of Michigan Press. Email: jmswales@umich.edu

Notes

* We thank Stefan Th. Gries for his advice on the multifactorial statistics, and Matthew Brook O'Donnell for his permission to use the n-gram data base he designed. Any remaining errors are entirely our own.

1. See ⟨http://micusp.elicorpora.info⟩.
2. As in MICASE, the Michigan Corpus of Academic Spoken English, we used the University of Michigan's Academic Division categories.
3. For more detailed information about MICUSP, its design and compilation, the reader is referred to Ädel and Römer (Forthcoming) and Römer and O'Donnell (2011) and O'Donnell and Römer (Forthcoming).
4. Sentence-initial cases of *this* were here defined as instances in which *this* is part of a main clause subject, the main clause potentially being preceded by adverbials, conjunctions, or quantifiers (*However, this . . . /And this . . . /All this . . .*), and potentially preceded by a subordinate clause. For the cluster analyses in Sections 4.3, only those instances of sentence-initial *this* in which *this* constitutes the first word of a new sentence were taken into consideration, amounting to a total of 4,200 instances.
5. Other applications of collostructional analysis include studies of dialectal variation (Wulff, Stefanowitsch and Gries 2007, Mukherjee and Gries 2009), diachronic stages (Hilpert 2006, Gries and Hilpert 2008), and accuracy in learner language (Gilquin forthcoming, Wulff and Gries 2011).
6. See Stefanowitsch and Gries (2003: 217–218) for justification of using the Fisher Yates exact test.

7. More precisely, the minimal adequate model will contain significant predictors and non-significant predictors as long as the latter are part of a significant interaction. We present a minimal adequate model based on our data sample in Section 4.2.
8. We computed the logistic regression in *R* using the functions glm to obtain logistic models, the function Anova(model.glm, type = "III", test.statistic = "Wald") for model comparisons, and lrm(formula = formula(model.glm), x = T, y = T, linear.predictors = T) to obtain a summary of the predictive power of the minimal adequate model as a whole as well as the predictive power of each predictor therein.
9. On the positional variation of phraseological items in MICUSP, see also O'Donnell and Römer (In preparation).
10. *This*-clusters of spans five and six have not been listed in Table 5 because with increasing span size, occurrence numbers drop significantly so that only the top two or three items in the cluster lists occur more than five times. Most frequent in the *Collocate* output lists are the five-word *this*-clusters *this is not to say* (23 hits) and *this is due to the* (10 hits), and the six-word *this*-clusters *this is not to say that* (21 hits) and *this is due to the fact* (5 hits), both extensions of four- and five-word clusters.
11. We decided to exclude verbs from the phraseological part of the analysis that are distinctively associated with unattended *this* but occur less than 20 times in a sentence-initial *this* + verb-cluster (e.g. *leave*, *increase*, *cause*) because with items of such comparatively low frequency, it tends to be difficult to reliably identify formal or functional patterns.

References

Ädel, Annelie & Ute Römer. (Forthcoming). Research on advanced student writing across disciplines and levels: Introducing the Michigan Corpus of Upper-level Student Papers. *International Journal of Corpus Linguistics*.

Axelrod, Rise B. & Charles R. Cooper. 2008. *The St Martin's guide to writing*, 8th edn. Boston: Bedford/St Martin's.

Barlow, Michael. 2004a. *Collocate 1.0: Locating collocations and terminology*. Houston, TX: Athelstan.

Barlow, Michael. 2004b. *MonoConc Pro 2.2 (MP2.2)*. Houston, TX: Athelstan.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow, UK: Pearson Education.

Charles, Maggie. 2003. 'This mystery . . .': A corpus-based study of the use of nouns to construct stance in theses from two contrasting disciplines. *Journal of English for Academic Purposes* 2. 313–326.

Csomay, Eniko. 2009. Positioning lexical bundles in discourse structure: The case of classroom teaching. Paper presented at the American Association for Applied Linguistics Conference, Denver, Colorado. 21–24 March.

Ede, Lisa. 2004. *Work in progress*. 6th edn. Boston: Bedford/St. Martins.

Faigley, Lester. 2007. *Writing: A guide for college and beyond*. New York: Pearson/Longman.

Finn, Seth. 1995. Measuring effective writing: Cloze procedure and anaphoric "this". *Written Communication* 12. 240–266.

Francis, Gill. 1986. *Anaphoric nouns*. Birmingham (UK): English Language Research.

Geisler, Cheryl, David S. Kaufer & Erwin R. Steinberg. 1985. The unattended anaphoric "this": When should writers use it? *Written Communication* 2. 129–155.

Gilquin, Gaëtanelle. Forthcoming. Lexical infelicity in causative constructions: Comparing native and learner constructions. In Jaakko Leino and Ruprecht von Waldenfels (eds.), *Analytical causatives*. Munich: Lincom Europa.

Gries, Stefan Th. 2003. *Multifactorial analysis in corpus linguistics: A study of particle placement*. London/New York: Continuum.

Gries, Stefan Th. 2007. Coll.analysis 3.2. A program for R for Windows 2.x.

Gries, Stefan Th. & Martin Hilpert. 2008. The identification of stages in diachronic data: Variability-based neighbor clustering. *Corpora* 3(1). 59–81.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9(1). 97–129.

Harrell, Frank E. Jr. 2001. *Regression modeling strategies. With applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Hilpert, Martin. 2006. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory* 2(2). 243–257.

Hoey, Michael P. 2005. *Lexical priming: A new theory of words and language*. London: Routledge.

Hoey, Michael P. 2009. Corpus-driven approaches to grammar: The search for common ground. In Ute Römer and Rainer Schulze (eds.), *Exploring the lexis-grammar interface*, 33–47. Amsterdam: John Benjamins.

Hoey, Michael P. & Matthew B. O'Donnell. 2008. Lexicography, grammar, and textual position. *International Journal of Lexicography* 21(3). 293–309.

Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: CUP.

Hyland, Ken. 1998. *Hedging in scientific research articles*. Amsterdam: John Benjamins.

Johnson-Sheehan, Richard. 2005. *Technical communication today*. New York: Pearson/Longman.

Jurafksy, Daniel. 2003. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Rens Bod, Jennifer Hay and Stefanie Jannedy (eds.), *Probabilistic linguistics*, 39–96. Cambridge, MA: MIT Press.

Keune, Karen, Mirjam Ernestus, Roger Van Hout & R. Harald Baayen. 2005. Social, geographical, and register variation in Dutch: From written MOGELIJK to spoken MOK. *Corpus Linguistics and Linguistic Theory* 1. 183–223.

Mahlberg, Michaela & Matthew B. O'Donnell. 2008. A fresh view of the structure of hard news stories. In Stella Neumann and Erich Steiner (eds.), *Online proceedings of the 19th European Systemic Functional Linguistics Conference and Workshop, Saarbrücken, 23–25 July 2007*. http://scidok.sulb.uni-saarland.de/volltexte/2008/1700/.

Markel, Mike. 2004. *Technical communication*, 7th edn. Boston: Bedford/St. Martins.

Mukherjee, Joybrato & Stefan Th. Gries. 2009. Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English. *English World-Wide* 30(1). 27–51.

O'Donnell, Matthew B. & Ute Römer. In preparation. Positional variation of n-grams and phrase-frames in a new corpus of proficient student writing.

O'Donnell, Matthew B. & Ute Römer. Forthcoming. From student hard drive to web corpus (Part 2): The annotation and online distribution of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. Harlow, UK: Longman.

Römer, Ute. 2010. Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction* 3(1). 95–119.

Römer, Ute & Matthew B. O'Donnell. 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(2). 159–177.

Römer, Ute & Stefanie Wulff. 2010. Applying corpus methods to writing research: Explorations of MICUSP. *Journal of Writing Research* 2(2). 99–127.

Sinclair, John. 2004. *Trust the text. Language, corpus and discourse*. London: Routledge.

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions. Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.
Strunck, William & E. B. White. 1979. *The elements of style*. London: Macmillan.
Swales, John M. 2004. *Research genres. Explorations and applications*. Cambridge: Cambridge University Press.
Swales, John M. 2005. Attended and unattended "this" in academic writing: A long and unfinished story. *ESP Malaysia* 11. 1–15.
Swales, John M. & Christine B. Feak. 2000. *English in today's research world: A writing guide*. Ann Arbor, MI: University of Michigan Press.
Swales, John & Christine B. Feak. 2004. *Academic writing for graduate students*, 2nd edn. Ann Arbor, MI: University of Michigan Press.
Williams, Joseph M. 1985. *Style: Ten lessons in clarity and grace*. (2nd ed.). Glenview, IL: Scott, Foresman.
Wulff, Stefanie. 2008. *Idiomaticity: A Usage-based Approach*. London/New York: Continuum.
Wulff, Stefanie & Stefan Th. Gries. 2011. Corpus-driven methods for assessing accuracy in learner production. In Peter Robinson (ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance*, 61–88. Amsterdam, Philadelphia: John Benjamins.
Wulff, Stefanie, Anatol Stefanowitsch & Stefan Th. Gries. 2007. Brutal Brits and persuasive Americans: Variety-specific meaning construction in the *into*-causative. In Günter Radden, Klaus-Michael Köpcke, Thomas Berg and Peter Siemund (eds.), *Aspects of meaning construction*, 265–281. Amsterdam, Philadelphia: John Benjamins.