Applying corpus methods to written academic texts: Explorations of MICUSP

Ute Römer° & Stefanie Wulff*

°University of Michigan, Ann Arbor - *University of North Texas, Denton | USA

Abstract: Based on explorations of the Michigan Corpus of Upper-level Student Papers (MICUSP), the present paper provides an introduction to the central techniques in corpus analysis, including the creation and examination of word lists, keyword lists, concordances, and cluster lists. It also presents a MICUSP-based case study of the demonstrative pronoun *this* and the distribution and use of its attended and unattended forms in different disciplinary subsets of the corpus. The paper aims to demonstrate how corpus linguistics and corpus methods can contribute to writing research and provide fruitful insights into student academic writing.

Keywords: MICUSP, student academic writing, corpus analysis, attended/unattended *this*, disciplinary variation

journal of WRITING RESEARCH

Römer, U. & Wulff, S. (2010). Applying corpus methods to writing research: Explorations of MICUSP. *Journal of Writing Research, 2*(2), 99-127.
Contact and copyright: Earli | Ute Römer, University of Michigan, English Language Institute, 500 E. Washington St., Ann Arbor, MI 48104-2028 | USA – uroemer@umich.edu. This article is published under *Creative Commons Attribution-Noncommercial-No Derivative Works 3.0* Unported license.

1. Introduction

Over the last two decades, corpus linguistics has started to turn from a pure methodology into a fully-fledged discipline. In fact, various theoretical concepts and frameworks such as Hunston and Francis' (2000) *Pattern Grammar* or Hoey's (2005) concept of *Lexical Priming* have emerged from corpus-linguistic approaches to language. Moreover, corpus linguistics has been shown to be particularly compatible with contemporary usage-based linguistic frameworks, including Cognitive Linguistics (Schönefeld, 1999), Construction Grammar (Goldberg, 2006), and Discourse Analysis (Baker, 2006). Likewise, corpus data are increasingly used as supplementary data in psycho-linguistic and first/second language acquisition research (Tomasello, 2003; Ellis and Larsen-Freeman, 2009).

Diverse as many of these frameworks and their thematic foci may be, they share the common assumptions that linguistic theorizing should be driven first and foremost by (representative samples of) authentic language data, and that a solid linguistic hypothesis and theoretical claims should be based on a thorough description of these data with regard to the phenomenon under investigation. As Tognini-Bonelli (2001) puts it, '[t]he theory has no independent existence from the evidence and that general methodological path is clear: observation leads to hypothesis leads to generalization leads to unification in theoretical statement.' (p. 84-85)

In other words, corpus linguistics can assist the researcher to assess and describe a linguistic phenomenon in a maximally objective and hence largely theory-neutral fashion. As such, corpus linguistics is fundamentally incompatible only with linguistic frameworks in which theoretical assumptions and hypotheses guide the analysis, which are then tested against the researcher's intuition.

Recent publications in corpus linguistics have also recognized writing as a field worthy of investigation, covering topics ranging from genre-analytical approaches to research articles (Hyland, 1998) to analyses of learner writing (Altenberg & Granger, 2001; Ädel, 2006) or the stylistics of thought representation (Semino & Short, 2004). Likewise, a 2006 special issue of *IEEE Transactions on Professional Communication* was devoted to the question what corpus linguistics can contribute to research in professional communication. In the introduction to this special issue, Orr (2006) notes:

> Corpus linguistics has much to offer the field of professional communication, for it allows researchers to study spoken or written discourse in considerable detail, which can yield information about language structure or use that is normally beyond the grasp of intuition and personal experience. By carefully designing corpora that are representative of language as it is actually being used today (or was used in the past) and then analyzing the data with proper methods and technologies, researchers can better understand a rather wide variety of things that might be of use to professional communicators as well as to those who support them. (p. 213)

Orr lists various examples that cover the breadth of the field of writing research (and professional communication more generally) and that could benefit from a corpus-linguistic perspective: the assessment of user-friendliness of online-help functions; the identification of discipline-specific core vocabulary for non-native college students; the contrastive analysis of language features associated with written peer- vs. public communication; the identification of successful résumés; or the identification of characteristics of problematic product assembly instructions (Orr, 2006, p. 213-214).

In spite of the growing recognition of the usefulness of corpus linguistics for professional communication research in general and writing research in particular, it is hard to find a basic introduction to corpus linguistic methods tailored to the needs of writing researchers. The present paper seeks to take a first step toward closing this gap and accordingly has two main objectives. Firstly, we would like to acquaint readers who may not be familiar with corpus work with the core techniques in corpus analysis. Secondly, we will illustrate, by means of a case study, how corpus tools can be employed to highlight important aspects of a text or text collection that may go unnoticed otherwise. In doing so, we hope to demonstrate the potential of corpus-analytic techniques for the field of writing research at large, be it as a primary method of investigation, or a supplementary method to test, complete, and qualify given assumptions.

Although a number of different linguistic subfields and theoretical frameworks rely on corpus analysis, there is still a shortage of certain types of corpora (which may not come as a surprise if we consider the variety of possible corpus types and the amount of time/money that goes into corpus compilation). For example, hardly any corpora representing proficient (native speaker) student writing, particularly at the graduate level, have been made publicly available to date. A recent attempt to fill this gap is the *Michigan Corpus of Upper-level Student Papers* (MICUSP) that was compiled at the University of Michigan English Language Institute and released to the public in 2009. Designed in analogy to its well-known sister corpus MICASE (the *Michigan Corpus of Academic Spoken English*), MICUSP covers a variety of academic disciplines from the Humanities, Social Sciences, Biological and Health Sciences, and Physical Sciences. It allows for complex searches by discipline, writer level, and writer characteristics (e.g. native-speaker status or sex). A minor objective of the present paper is to introduce MICUSP as a resource for the investigation of upper-level student writing across disciplines (see Section 2). In Section 3 of this paper, we offer a hands-on tutorial on corpus analysis using the freeware package *AntConc*. We illustrate how corpus methodology, including the analysis of (key)word lists, concordances, and word clusters, can provide fruitful insights into writing, in particular student academic writing.

In Section 4 we then present a MICUSP-based case study that addresses a prominent topic in writing research and teaching: attended vs. unattended *this*. The case study revisits the variable realization of the demonstrative pronoun *this* attended by a noun or noun phrase, as in *This behavior may also be due to the materials non-linearity*, or

standing alone, as in *This may have implications for instructors who want students to produce academic text* (examples taken from MICUSP). Our analyses of more than 9,000 instances of *this* in a pre-release version of MICUSP show that a corpus approach can uncover aspects of the distribution, function and use of language features that would most likely go unnoticed by non-corpus approaches. We will round off the paper with a summary of our findings and some concluding thoughts on the potential of corpus analysis in researching writing.

2. The Michigan Corpus of Upper-level Student Papers (MICUSP)

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 (see http://micusp.elicorpora.info). The corpus, the first of its kind in North America, enables corpus researchers, EAP teachers, and testers to investigate the written discourse of highly 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 late 2009.[1]

MICUSP consists of 829 papers (totalling around 2.6 million words) of different types (e.g. essays, reports, response papers) from altogether 16 different disciplines within four subject divisions (Humanities and Arts, Social Sciences, Biological and Health Sciences, and Physical Sciences). All papers included in MICUSP were written by final year undergraduate and first to third year graduate students who obtained an A grade for their paper. Each of the papers in MICUSP has been marked up in 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 sex, 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 January 2009, henceforth MICUSP_Jan09. This version of the corpus consists of 623 A-graded student papers from 16 different disciplines, including Biology, Education, English, Linguistics, Mechanical Engineering, Nursing, Physics, and Sociology. The 623 texts in MICUSP_Jan09 make up approximately 1.25 million words. File headers, titles, abstracts, references, and appendices have been excluded for this version, which consists of body text sections only. The files have been organized into subsets according to discipline and student level so that targeted searches can be performed and search results can be reported separately for groups of papers.

3. Central steps in corpus analysis

Let us now examine how a corpus like MICUSP_Jan09 can be accessed by the researcher, writing instructor or student, and how useful information can be retrieved

from it. A number of available software tools, so-called 'concordance programs' or 'concordancers', enable easy electronic access to the texts stored in a corpus and provide a range of functions to analyze language phenomena and highlight interesting aspects about the language captured in the corpus. Three of the most commonly used software packages for corpus analysis are *WordSmith Tools*, *MonoConc Pro*, and *AntConc*. While the first two packages are commercial and require a license, *AntConc* is free, which is one of the reasons why we decided to feature it in this article. Without a concordance program like *AntConc*, a corpus would be of no use other than being an electronic repository of texts that could then be read on screen (or on paper printouts) in the normal linear fashion. The concordancer, however, allows different (and faster) ways of accessing corpus texts. Basically, what the software does is it "selects, sorts, matches, counts and calculates" (Hunston & Francis, 2000, p. 15). In doing so, it provides different views on the data captured in the corpus, e.g. it may highlight what the most frequent 3-word combination is or which words tend to occur immediately to the left of the noun *problem* in a certain type of discourse. Following Barlow (2004, p. 205), we will regard text or corpus analysis as text or corpus transformation and show in what ways different types of transformation can draw attention to different aspects of a text or corpus. In this paper, we will provide a step-by-step introduction to some core corpus analytic (or text transformational) techniques using *AntConc*. These include the creation of a word list and keyword list (see 3.1), compiling and analysing a concordance (3.2), tracing repeated instances of a word or phrase in a text (3.3), and examining contextual phenomena such as collocates and clusters (3.4).

AntConc was developed by Laurence Anthony of Waseda University, Tokyo, Japan (see Anthony, 2006), originally for use in the technical writing classroom. The software is free for download from the author's homepage.[2] There are versions for different platforms available (Windows, Macintosh, Linux). *AntConc* is sporadically updated; the version used in this paper is *AntConc* 3.2.1. Information about the program and its tools can be found in the Readme file on Anthony's website. *AntConc* does not require any installation on your computer but can be launched by simply double-clicking on the executable file (in our case 'antconc3.2.1w.exe'). Once you have started the program, the screen displayed in Figure 1 appears. It shows a small frame on the left which, once a corpus has been loaded, gives a list of files and a larger frame with seven tabs, one for each tool.

Before you can perform any of the actions described below, you need to select a text or corpus to base your analyses on. To load texts, go to the *AntConc* 'File' menu and use either the 'Open File(s)...' or the 'Open Dir...' option (if your files are in a number of subfolders, the latter option is a time-saver). The list of selected files will be displayed in the left column of the *AntConc* window under 'Corpus Files'. The MICUSP_Jan09 files we loaded are all in plain text (txt) format; it is also possible to load data in xml or html format.

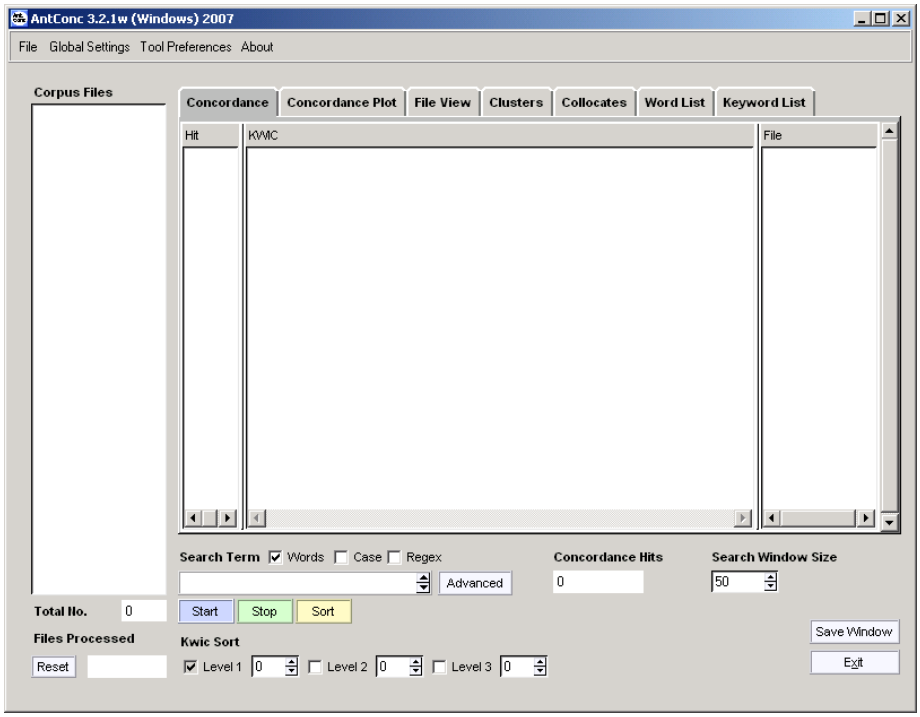

Figure 1. Top screen of *AntConc* (version 3.2.1 for Windows) with Concordance tool selected (prior to loading corpus files)

3.1 Creating a word list and keyword list

A useful first step in approaching a corpus or text is to generate a list of all the words that occur in it together with their frequencies. The generation of a word list is also, as Barlow (2004) notes, "[p]robably the most radical transformation of a text used in linguistic analysis" (p. 207). Word lists are useful because they highlight which words are most frequent in a corpus and may be worth investigating. The 'Word List' tool in *AntConc* is able to create alphabetical and frequency-sorted lists from the file(s) loaded. The tool also offers reverse ordering, ordering according to word endings (grouping words together that end in *-a*, *-ca*, *-ica*, and so on), and case-sensitive word listing. Once compiled, a word list can be resorted and saved to a text file using the 'Save Output to Text File...' command from the 'File' menu, or to a new window by means of the 'Save Window' button. A saved text file can then be opened and edited in a spreadsheet software or in any standard text editor.

Table 1 displays the top-20 items in a frequency-sorted word list based on MICUSP_Jan09. As we can see, all items in the list are function words and do not tell us much about what the texts in the corpus are about. This does not mean, however, that a closer look at selected high-frequency function words may not lead to interesting

observations. It may be worth investigating, for example, why the demonstrative pronouns *that* and *this* are so highly frequent in our advanced student papers. If we scroll further down the list, we see that the most frequent nouns in MICUSP_Jan09 are *students*, *time*, *people*, *system*, *study*, and *data* – all important items in academic discourse and potential starting points for analyses of student writing.

Table 1. Top-20 items in a frequency-sorted MICUSP_Jan09 word list

Rank	Frequency	Word
1	81,410	the
2	46,380	of
3	36,441	and
4	36,197	to
5	28,363	in
6	24,571	a
7	18,061	that
8	17,910	is
9	11,938	for
10	11,209	as
11	9,411	this
12	8,960	be
13	8,695	with
14	8,205	s
15	8,160	are
16	7,971	it
17	7,337	on
18	6,921	not
19	6,311	by
20	5,457	From

While a word list highlights what is frequent in a corpus or text, it does not tell us what is important or unusually frequent. To identify the most outstanding or unexpectedly frequent words, *AntConc* offers a 'Keyword List' tool that compares a frequency wordlist based on the corpus under analysis (your target corpus) with another frequency wordlist based on a reference corpus (usually a larger corpus of a more general type). The tool then lists outstanding words in order of their 'keyness' values. Words get a high keyness value if they occur considerably more frequently in a selected corpus than they would be expected to occur on the basis of figures derived from a reference corpus. To create a keyword list, you need to select a target corpus to perform the keyword extraction on and go to the 'Keyword List' settings in the 'Tool Preferences' menu. In the settings, choose a reference corpus in the same way you selected the target corpus and click 'Apply'. You then go to the 'Keyword List' tab and press the 'Start' button (if you have not created a word list from your target corpus, *AntConc* will at this point inform you that it needs to jump to the Word List tool).

Table 2. Top-20 keywords in the Biology subsection of MICUSP_Jan09 (reference corpus: MICUSP_Jan09)

Rank	Frequency	Keyness	Keyword
1	614	839.352	et
2	608	827.795	al
3	356	695.874	species
4	263	567.877	genes
5	183	379.511	gene
6	143	317.024	plague
7	163	304.576	cells
8	137	293.282	protein
9	171	289.929	females
10	123	282.758	leptin
11	124	281.478	mutations
12	177	281.282	males
13	121	276.959	crosses
14	118	267.687	mutant
15	136	253.891	host
16	176	241.314	color
17	133	241.127	selection
18	121	235.659	genetic
19	122	235.117	eye
20	103	234.39	flies

Table 2 shows the top-20 items in a keyword list based on the Biology subsection of MICUSP_Jan09 (64 papers written by students in Biology), with the whole MICUSP_Jan09 used as reference corpus. As we can see here, a keyword analysis clearly highlights academic expressions (the two top items are *et* and *al*.) and discipline-specific vocabulary. Words like *species*, *gene(s)*, *plague*, *cells*, and *protein* obtain highest keyness values and indicate what the texts covered in the subcorpus are about. This demonstrates that a keyword list can be a useful tool in the disciplinary writing classroom because it highlights items that are important in a certain discipline and that students need to know. If we scroll down to the very end of our keyword list in *AntConc*, we see words highlighted in blue. These are 'negative keywords', i.e. words that occur comparatively more often in our reference corpus than in our target corpus and are negatively key in the target corpus (and have low keyness values). Negative keywords in our Biology subcorpus are, for example, *perceive*, *fields*, *organizational*, *metaphor*, and *governments* – words that are rare in Biology papers but common in other disciplinary subsets of MICUSP_Jan09.

3.2 Compiling and analysing a concordance

In the next analytic step, we are moving to the core tool in corpus linguistics: the concordance. Having torn the corpus texts apart in the creation of a word list and keyword list, we will now reverse the process and provide a contextualized view of

select items in our corpus, items that we would like to know more about (perhaps one or two of the particularly frequent or particularly key words). Barnbrook (1996) describes the main purpose of a concordance as follows: "The concordance provides a simple way of placing each word back in its original context, so that the *details of its use and behaviour* can be properly examined." (p. 65) (emphasis added) Concordances are usually displayed in KWIC (key word in context) format, with the search word (or phrase) shown in the middle of the screen and some context left and right of it. They list all instances of a word (or phrase) found in the selected corpus which saves us from going through each text file separately to pull out relevant examples.

Creating a concordance in *AntConc* is very straightforward. You select the 'Concordance' tab in the top screen, enter a search word (or phrase) in the box underneath the main window, and click the 'Start' button. Part of the concordance of the word *gene* (a keyword) in the MICUSP_Jan09 Biology subcorpus is shown in Figure 2. The 'Concordance Hits' box tells us that there are altogether 186 occurrences of *gene* in the 64 corpus files we loaded. We also see in which of the 64 files each of the hits occurs ('File' column on the right of the 'KWIC' display). The value of 50 in the 'Search Window Size' box refers to the number of characters to be displayed on either side of the search word. By default, *AntConc* concordance searches are case insensitive but they can also be made case sensitive by ticking the 'Case' box next to 'Search Term'. A case-sensitive search can, for instance, help distinguish between sentence-initial *However* (1,102 hits in MICUSP_Jan09) and non-sentence-initial *however* (677 hits in MICUSP_Jan09).

Since the size of the computer screen (and the *AntConc* window) is limited, only a certain amount of context can be displayed in each concordance line. Depending on the type of analysis, it may be necessary to look at more context for some of the search words (e.g., in a search for sentence-initial *However*, we may want to read the sentence that precedes *However*). The *AntConc* 'File View' tool makes it possible to view any of the loaded files at any time. You can go to a file either by clicking on the search word in the concordance line you would like to expand (the cursor changes to a small hand icon when you move it over the search word), or by clicking on the 'File View' tab and then on any of the file names in the 'Corpus Files' list in the left-hand window. Figure 3 presents the file view for the first line in our *gene* concordance search (with *gene* marked in black). To carry out a new search in the selected file, you just need to type a word or phrase in the 'Search Term' box and hit 'Start'.

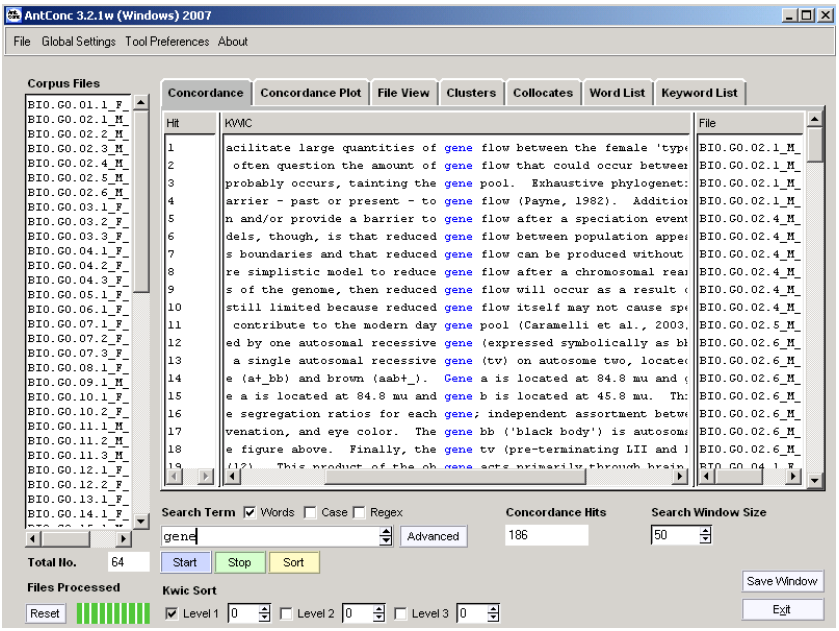

Figure 2. *AntConc* concordance of the word *gene* in the MICUSP_Jan09 Biology subcorpus

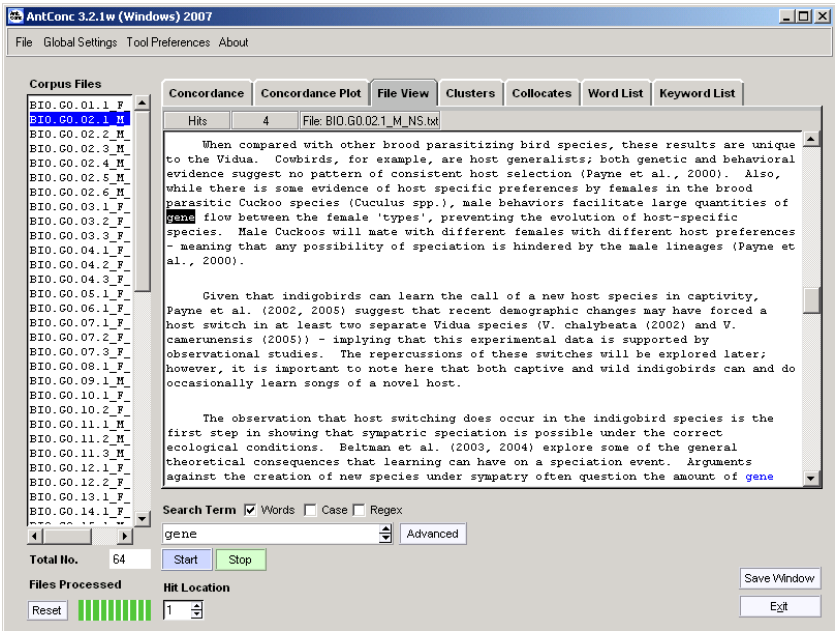

Figure 3. *AntConc* 'File View' display for the word *gene* in MICUSP_Jan09 file BIO.G0.02.1_M_NS.txt

Unlike a text that we usually read horizontally, line by line, a concordance is read vertically, focussing on the search word (or 'node') in the middle of the screen (cf. Tognini-Bonelli, 2001, p. 3). By looking at the context words on the left and on the right of the node, you get access to phraseological patterns and to the meanings expressed by the search word or phrase. What you search for in the concordance are repeated events – repetitions of words in combination with other words. Sorting the context in a concordance facilitates the identification of repeated events and makes patterns visible. If you sort a concordance, e.g. the *gene* concordance displayed in Figure 2, the order of the concordance lines is rearranged according to certain predefined sorting criteria and lines that contain the same words to the left or right of the search word or phrase are grouped together. In *AntConc*, you can sort a concordance by selecting levels (positions to the left or right of the search word) under 'Kwic Sort' and pressing the 'Sort' button. Figure 4 shows part of the *gene* concordance with the context words sorted alphabetically to the right by three positions (1R, 2R, 3R).

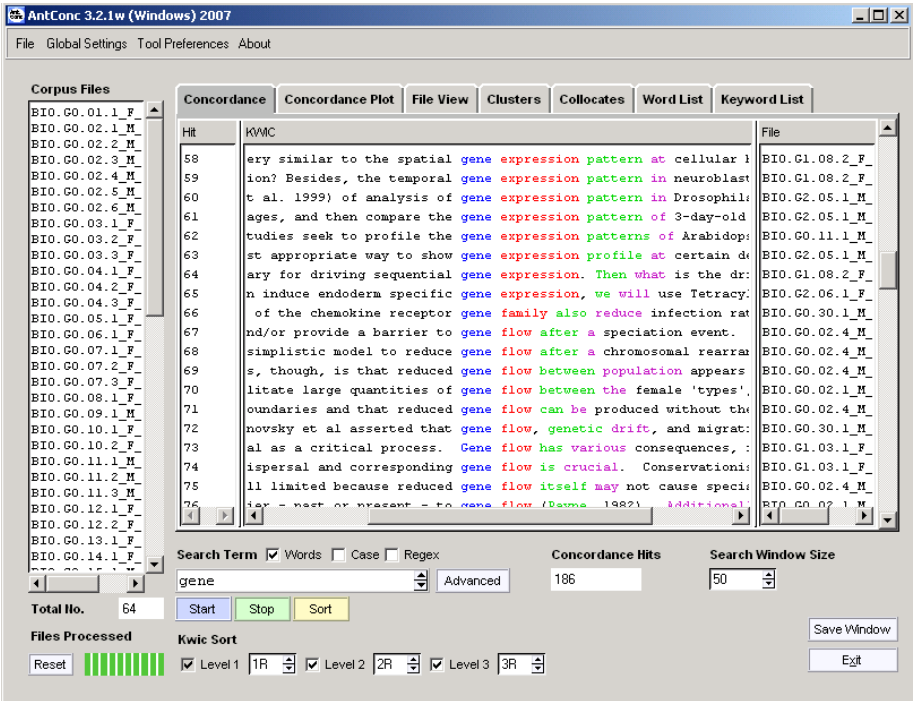

Figure 4. Part of a right-sorted *AntConc* concordance of the word *gene* in the MICUSP_Jan09 Biology subcorpus

The right-sorted concordance highlights terms such as *gene expression pattern*, *gene flow*, or *gene promoter* (further down the list). If the same concordance is sorted alphabetically to the left (1L, 2L, 3L), different potentially interesting word combinations are highlighted, e.g. *body color gene*, *mutant gene*, *wing venation gene*, and *endoderm expressive gene* (followed by *promoter* or *expression*). We will deal with more options *AntConc* offers to highlight repeated combinations of words in a corpus or text in Section 3.4 below.

3.3 Tracing repeated instances of a word or phrase in a text

Sometimes it may be useful to know not only *what* the most common (or the most unusually common) words in a text or corpus are and how they combine with other words but also *where* they occur in a text and how evenly they are distributed across different texts in a corpus. You may, for example, want to find out whether *gene* occurs in all 64 MICUSP_Jan09 Biology papers or whether it is used frequently in some papers but rarely or not at all in others. Also, it may be interesting to know if a selected word has a preference to occur at the beginning or end of a text, if it clusters in a certain section of a text or is evenly distributed across a text.

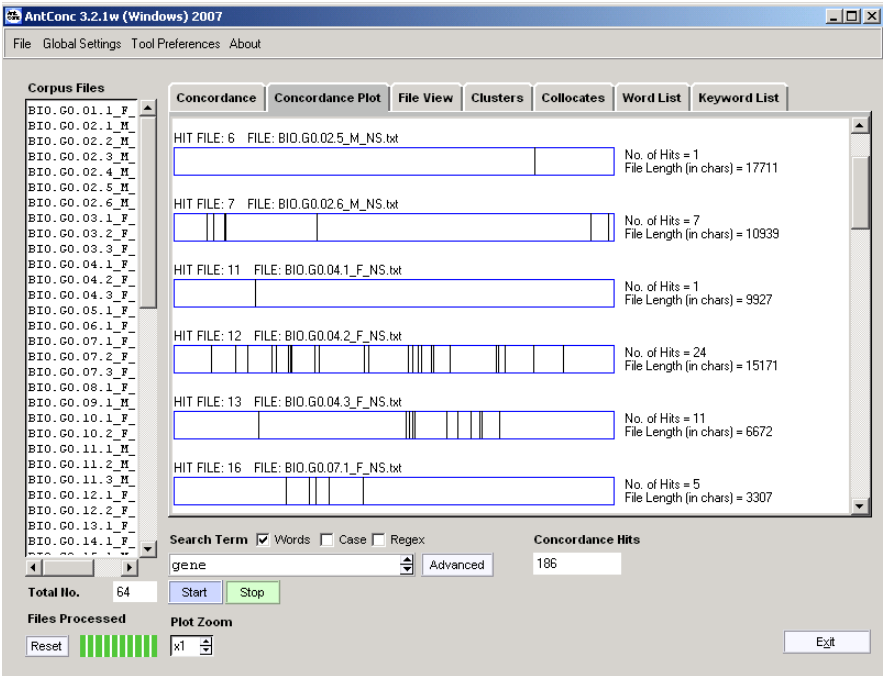

Figure 5. *AntConc* concordance plot for the word *gene* in the MICUSP_Jan09 Biology subcorpus

AntConc offers a tool that serves to visualize repeated instances of a word or phrase in a text: the 'Concordance Plot' tool. In a concordance plot, all instances of a word are visualized in the form of a barcode, separately for each corpus file. Each line in the barcode represents an occurrence of the search word in a text. Part of the concordance plot for the distribution of the word *gene* across the MICUSP_Jan09 Biology files is given in Figure 5. The total of 30 barcodes indicates that *gene* occurs in 30 of 64 files and is not evenly distributed across texts. The number of hits per file ranges from 1 to 24, and in some cases we find that the word clusters in a certain part of the text, e.g. in the second quarter in file BIO.G0.07.1_F_NS.txt. A click on a line in any of the barcodes takes you to the file view with the search word (here *gene*) highlighted in the text.

3.4 Examining contextual phenomena: Collocates and clusters

In Section 3.2 we dealt with the concordance as the central tool in corpus analysis and discussed how sorting the context in a concordance can help highlight patterns in texts (see the sorted *gene* concordance in Figure 4). Corpus analysis offers, however, other means to uncover patterns or repeated phrases and word associations in texts. We can examine how words collocate (i.e. how they commonly co-occur with each other) and how they form word combinations or word clusters. *AntConc* provides us with two tools to investigate patterns and contextual phenomena: the 'Collocates' tool and the 'Clusters' tool.

Starting from a concordance search, the 'Collocates' tool generates a list of words that frequently occur in the context of the search word or phrase in the selected corpus files. The user selects the contextual span for the search, i.e. the window of words to both sides of the search word in which to find collocates. A span commonly used in corpus analysis is 5L to 5R (five words to the left and right). The *AntConc* collocates listing includes the frequencies of co-occurring words on the left ('Freq(L)') and on the right ('Freq(R)') of the search word as well as an optional statistical measure (Mutual Information or t-score; 'Stat' column). Figure 6 displays the top of a MICUSP_Jan09 collocates list for the word *interesting*, sorted by Mutual Information values. As we can see, the most significant collocates of *interesting* in our corpus are *note*, *applications*, *phenomena*, *approaches*, and *particularly*, as in *interesting to note*, *interesting applications*, *interesting phenomena*, *interesting approaches*, and *particularly interesting* – a finding that tells us something about the word combinations that are commonly used and the meanings that are created in student academic papers.

The *AntConc* 'Clusters' tool also provides insights into word patterning. It extracts clusters around a specified search word from a corpus and displays them together with their frequencies of occurrence. Clusters are word sequences of a pre-defined size or length, and clusters of different lengths can be extracted in a single step by entering different numbers in the 'Min. Size' and 'Max. Size' boxes. A MICUSP_Jan09-based search for *interesting* clusters of two to five words in length resulted in the list shown in Figure 7.

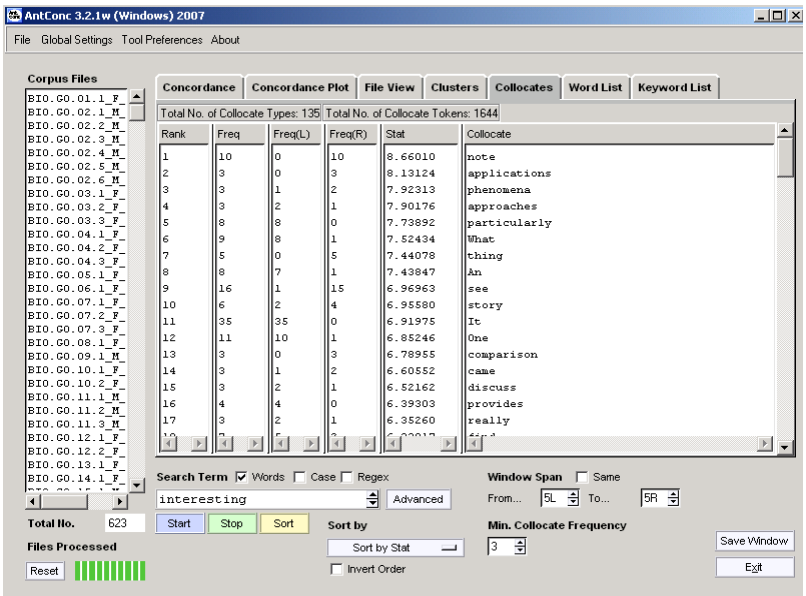

Figure 6. *AntConc* collocates list for the word *interesting* in MICUSP_Jan09, sorted by statistical measure (Mutual Information)

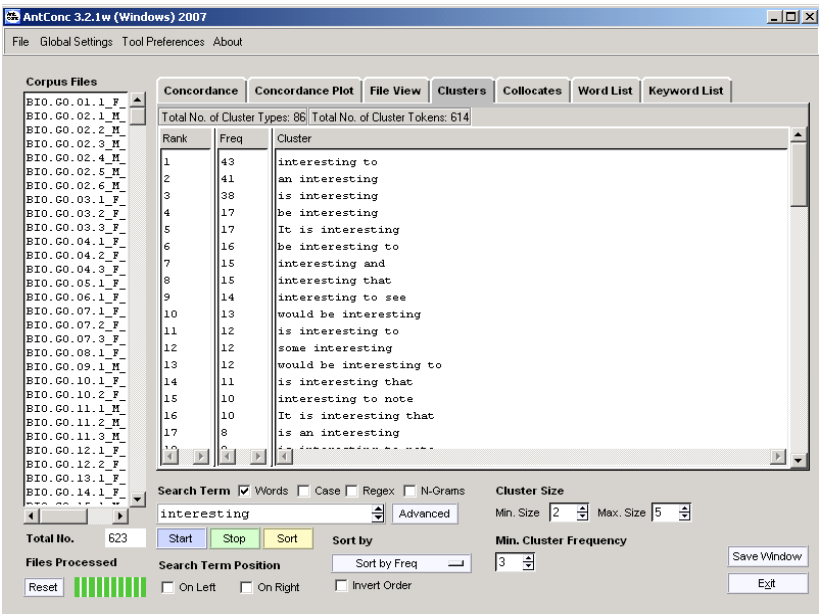

Figure 7. *AntConc* clusters list for the word *interesting* in MICUSP_Jan09, cluster size: 2-5

The list indicates that student writers (across disciplines) commonly use *interesting* in phrases like *it is interesting*, *interesting to see*, *would be interesting*, and *interesting to note*, which tells us something about the ways in which they structure the discourse and express evaluation.

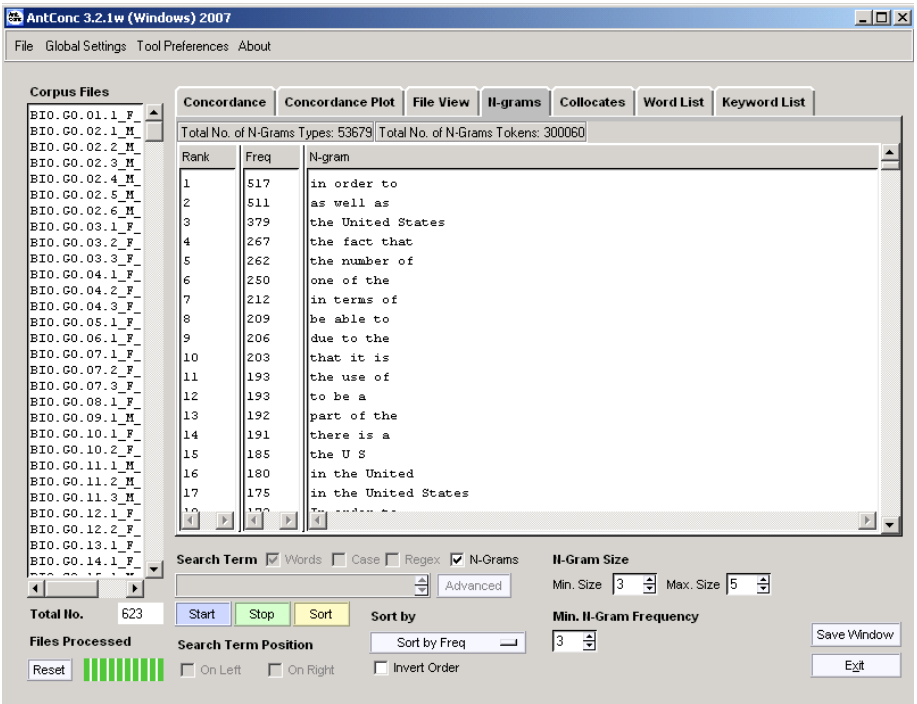

Figure 8. Frequency-sorted list of n-grams (size: 3-5) in MICUSP_Jan09

Within the 'Clusters' tool, it is also possible to extract word clusters of varying lengths from a corpus without specifying a search word. This can be done by activating the 'N-Grams' box underneath the main search window and determining the minimum and maximum n-gram size. *AntConc* then creates a list of all combinations of n words, e.g. 3-grams, 4-grams and 5-grams, that occur repeatedly in a corpus. As Figure 8 shows, the 3-grams *in order to*, *as well as*, *the fact that*, *the number of*, *in terms of*, and *due to the* are among the most frequent word sequences in MICUSP_Jan09 and hence may deserve special attention in studying and teaching academic writing. By clicking on an item in a clusters or n-grams list, it is possible to go directly to a KWIC concordance display of the selected item and examine the cluster or n-gram in context.

In the previous sections we have discussed some central corpus-analytic techniques and applied them to MICUSP_Jan09. We have shown how a concordance package like

AntConc can provide a number of exciting views on the data captured in a corpus. To further demonstrate the potential of a corpus approach to investigating writing, we will now turn to a case study of attended and unattended *this* in MICUSP_Jan09.

4. Case study: Attended and unattended *this* in student writing

As we are not writing researchers but corpus linguists, we looked in the literature for a topic that would furnish us with a case study. We noticed that Swales and Feak (2004) and Swales (2005) paid considerable attention to the role of *this* as a common cohesive device in academic writing. Especially in sentence-initial position, the demonstrative, sometimes followed by a noun or noun phrase and sometimes not, is a key exponent of given-new information structuring. As Swales notes, this pattern is one clear way of "getting out of one sentence and into another", and hence has relevance for writing instructors and writing textbook authors. For these reasons, we decided to explore the occurrence of *this* + or - attendant noun in MICUSP.

4.1 Previous studies on *this* in academic writing

In English, *this* can either function as a demonstrative pronoun, as in (1), or a demonstrative determiner, as in (2) (unless indicated otherwise, examples are taken from MICUSP_Jan09).

1. <u>This</u> will raise the standard of living of the Americans as they can now afford to purchase a greater variety of goods and services.
2. <u>This change</u> is enough to transform the entire female low-skill labor market.

In (2), *this* is immediately followed by a noun phrase and accordingly often referred to as "attended" (Geisler et al., 1985) or as having an "associated nominal" (Huckin & Olsen, 1991). In analogy, the example of *this* in (1) can be described as "unattended". Writers' choice of one over the other variant can be described as a choice between the Gricean Maxim of Quantity (in the case of unattended *this*) and that of Manner (because attended *this* clarifies reference unambiguously) (Grice, 1975). The latter tends to be preferred by writing professionals not only for its clarity but also because it provides an opportunity for "higher-level recontextualization of the previous text" (Swales, 2005, p. 3), or to express interpretative stance – ergo its general association with a more professional style of writing.

When writers opt for attended *this*, the question arises which noun to select. One factor determining the variable presence of *this* seems to be the concreteness of the noun phrase. Swales (2005, p. 3) provides the following examples where the noun phrase in (3b) is comparatively more concrete than the one in (3c).[3]

3. a. Each chapter ends with a summary of the main points.
3. b. <u>This summary</u> is designed to help students studying on their own.
3. c. <u>This strategy</u> is designed to help students studying on their own.

Swales (2005) based his analysis on a subset of 80 research articles from eight academic disciplines taken from the Hyland corpus (Hyland, 1998). An analysis of the 50 most frequent noun phrases attending *this* in Hyland reveals that "there is a fair degree of convergence among many of the disciplines, except for philosophy and physics, which appear to have their own preferences" (Swales, 2005, p. 11). The biggest coherent groups of nouns appear to be metadiscoursal noun phrases (*study*, *article*, *paper*, *account*) and nouns relating to methodology (*method*, *technique*, *procedure*, *process*). Table 3 below provides an overview.

Table 3. Most frequent nouns attending *this* in the Hyland corpus (adapted from Swales, 2005, 10)

Dentistry	study (76)	finding (15)	result (5)	patient (5)	process (5)
Medicine	study (66)	group (8)	difference (7)	procedure (5)	technique (4)
Biology	result (14)	observation (7)	study (6)	difference (5)	finding (5)
Electrical Engineering	approach (14)	algorithm (11)	method (10)	paper (8)	technique (5)
Mechanical Engineering	paper (17)	method (8)	approach (7)	type (7)	figure (6)
Applied Linguistics	study (47)	result (10)	experiment (9)	difference (8)	finding (8)
Marketing	study (31)	paper (22)	cluster (13)	approach (12)	research (12)
Philosophy	account (10)	article (8)	argument (6)	conclusion (6)	claim (5)
Sociology	article (15)	model (10)	paper (10)	process (9)	group (6)
Physics	effect (9)	approach (7)	behavior (5)	contribution (5)	figure (5)

With regard to the choice between attended and unattended *this* in the first place, Celce-Murcia and Larsen-Freeman (1999) point out that the "demonstrative usage might be quite genre specific in written discourse" (p. 308). However, in accord with his analysis of the most frequent noun phrases across the academic disciplines covered in the Hyland corpus, Swales (2005, p. 10) found that apart from exceptionally low occurrences of attended *this* in Philosophy articles (44%) and comparatively high shares in the life/health sciences (75%), percentages of attended *this* as opposed to the unattended variant averaged around 64%.

4.2 Data Retrieval and Coding

We were interested to find out if, and to what extent, the variable presence of *this* is more tightly linked to discipline-specific tendencies in student writing compared to the corpus of published writing investigated by Swales (2005). In what follows, we will first look at the distribution of attended and unattended *this* across the 16 disciplines covered in MICUSP_Jan09 and then consider the head noun phrases attending all occurrences of attended *this* by discipline, as well as their dispersion across the disciplines.

Before we delve into the results of our case study, a brief note regarding methodology is in order. While the data retrieval for this case study was done using *AntConc*, the case study goes beyond the basic functions available in this software package. This is not indicative of any shortcomings of *AntConc*; rather, as we explain in more detail below, our research objectives required a thorough manual investigation of the concordance lines, which no concordance program will do on the researcher's behalf. To this end, we loaded the corpus into *AntConc* and did a simple search for *this*, retrieving a total of 9,411 hits. We then sorted the resulting concordance to the right (as explained in Section 3.2 above) and copied the concordance into a spreadsheet. The file names, which *AntConc* lists right of each concordance line in a separate window (see Figure 2), were also copied into this spreadsheet as a separate column. For all 9,411 hits, we then determined

- whether *this* was attended or not (this information required manual scanning, which was extremely facilitated by having the concordance sorted according to the right-hand context because, for instance, all "this is", "this was", or "this should" sequences are listed in one block and could immediately be coded as 'unattended'; similarly, nouns frequently following this, such as *study*, were listed together, so these instances could be coded quickly as 'attended');
- in which discipline *this* occurred (this information can be retrieved from the file names);
- provided that *this* was attended by a noun phrase, the head noun of that noun phrase (this information required manual combing of the data because the head of the noun phrase need not be the first word on the right of *this*; for instance, consider the complex noun phrase *this very interesting study*, where the head noun *study* occurs only in R3 position).

4.3 Results

First of all, we found that 6,839 (72.67%) out of 9,411 occurrences of *this* are attended, 2,572 (27.33%) are unattended. In other words, our data confirm that unattended *this* is clearly not a rare phenomenon at all, but constitutes a good share of all occurrences of *this*.

4.3.1. Distribution of (un)attended *this* by discipline

Let us now turn to the distribution of attended and unattended *this* across the 16 MICUSP_Jan09 disciplines; Table 4 provides an overview with absolute frequencies (n) and corresponding percentages (by discipline).

Table 4. Distribution of (un)attended *this* by discipline in MICUSP_Jan09

Discipline	Attended *this*		Unattended *this*		Total
	N	%	n	%	
Biology	631	77.42	184	22.58	815
Civil & Environmental Engineering	284	80.91	67	19.09	351
Classical Studies	150	72.46	57	27.54	207
Economics	325	79.08	86	20.92	411
Education	423	72.18	163	27.82	586
English	937	76.06	295	23.94	1,232
Industrial & Operations Engineering	385	73.47	139	26.53	524
Linguistics	484	71.70	191	28.30	675
Mechanical Engineering	185	76.45	57	23.55	242
Natural Resources & Environment	317	70.13	135	29.87	452
Nursing	533	70.60	222	29.40	755
Philosophy	476	62.14	**290**	**37.86**	766
Physics	67	69.79	29	30.21	96
Political Science	685	71.80	269	28.20	954
Psychology	679	70.51	284	29.49	963
Sociology	278	72.77	104	27.23	382
Total	6,839		2,572		9,411

Looking at Table 4, we note that the average percentage of attended *this* (73%) is substantially higher than the 64% average reported by Swales (2005). This may be a reflection of the slightly different disciplinary mix in the two corpora. Alternatively, one could argue that an increased use of attended *this* is a feature of less proficient writing, such that it functions as a cohesive device for lack of more sophisticated alternatives. If this were the case, we would expect to find some differences between student writing at different levels: the more experience students gain with academic writing, the lower the share of attended *this* should be. Fortunately, every text in MICUSP_Jan09 is annotated with information about the level of the student, so we can quickly check this hypothesis. Table 5 provides an overview of the distribution of (un)attended *this* by student level.

Table 5. Distribution of (un)attended *this* in MICUSP_Jan09 by student level

Student level	Attended *this*		Unattended *this*		Total
	n	%	N	%	
Final year undergraduate	3,927	71.23	1,586	28.77	5,513
First year graduate	1,300	74.03	456	25.97	1,756
Second year graduate	918	75.31	301	24.69	1,219
Third year graduate	694	75.19	229	24.81	923
Total	6,839		2,572		9,411

Table 5 is significant overall (χ^2=14.60; df=3; p<.002**): while final year undergraduate students produce more cases of unattended *this* than their more advanced peers, overall, the ratio of 3:1 for attended and unattended *this* is very stable across all four levels. In sum, the data do not suggest a development from higher to lower shares of attended *this*. A potential motivation for this difference between our student writers and the expert writers in the Hyland corpus could be that students feel more inclined to stick to the prescriptive grammar rule not to use unattended *this*. The more frequent omission of nouns or noun phrases following *this* in the texts captured in Hyland (i.e. published research articles) could also be related to word limits that the authors of the research articles had to stick to.

Returning to Table 4, we also find Swales's observation confirmed that attended *this* tends to be less frequent in Philosophy texts: only 62.14% of *this* are attended in the Philosophy subsection, the lowest percentage by far in our data (yet significantly higher than the 44% average that Swales observed for Philosophy papers written by established academics – which, again, may well reflect the impact of standards of 'proper' academic writing taught in many writing classes).

While the overall distribution of the table is highly significant (χ2=98.192; df=15; p<.001***), a closer look actually reveals that it is only the frequency of unattended *this* in Philosophy papers (highlighted in bold print in Table 4) that is responsible for the overall significance: it is considerably higher than we would expect (contribution to χ2=37.16; df=15; p<.001**). Beyond that, there are no major discipline-specific deviations from the general distribution of attended and unattended *this*. The less statistically inclined reader may wonder why the chi-square test did not pick up on the differences in frequency in, say, Biology (77% vs. 23%) as opposed to Physics (70% vs. 30%). However, it has to be kept in mind that percentages mask how many instances they are based on: In the Biology subsection of the corpus, we find a total of 815 instances of (un)attended *this*; the percentages for the Physics section, in contrast, are based only on 96 occurrences of (un)attended *this*. The chi-square test takes these differences into account and weighs the observed distributions accordingly.

On the other hand, while statistical significance tests are invaluable tools to quantify strong associations in the data, we would miss out on a number of interesting

tendencies by discarding the results as irrelevant on the basis of the failure of the data to meet an arbitrary significance threshold. At the same time, we may want to be able to compare the observed distribution across disciplines without having to bear in mind that the total number of occurrences of (un)attended *this* varies quite considerably across disciplines. Moreover, the different subsections of MICUSP_Jan09 representing these disciplines differ in size to begin with, as is summarized in Table 6.

Table 6. Numbers of words by academic discipline in MICUSP_Jan09

Discipline	Number of words
Biology	121,190
Civil & Environmental Engineering	40,249
Classical Studies	22,690
Economics	49,495
Education	81,301
English	180,016
Industrial & Operations Engineering	68,309
Linguistics	84,672
Mechanical Engineering	33,560
Natural Resources & Environment	68,038
Nursing	97,651
Philosophy	70,801
Physics	12,741
Political Science	147,651
Psychology	128,103
Sociology	49,807
Total	1,256,274

For instance, the Biology subsection comprises 121,190 words, and we find 815 hits of (un)attended *this* in this subsection; the Physics subsection, on the contrary, comprises only 12,741 words. Given the rather small size of the Physics subsection, the fact that we found only 96 occurrences of (un)attended *this* is no longer surprising, and similarly, we need to be careful how much weight we want to attribute to the observed distribution of attended and unattended *this*.

In order to facilitate comparing the observed distributions across subsections of a corpus that are of different size, corpus linguists often report relative and/or normalized frequencies rather than just absolute frequencies as given in Table 4. The relative frequency of, say, attended *this* in the Biology section can be obtained by dividing the number of occurrences of attended *this* (631) by the total number of words in the Biology section (121,190). Since the resulting number (631/121,190=.005) is small and hard to interpret (let alone compare with the number for other disciplines), we can additionally norm that number by an arbitrary value. Depending on the frequency of the phenomenon in question and the overall corpus (section) size, relative frequencies are typically normalized to ten thousand, a hundred thousand, or a million words. Sticking to our example, we can multiply the relative frequency of attended *this* in the Biology subsection by 10,000 to obtain a relative normalized frequency of 52. In other

words, attended *this* occurs on average 52 times in every 10,000 words in the Biology subsection of MICUSP_Jan09.

Accordingly, Table 7 provides the relative normalized frequencies for (un)attended *this* across all disciplines. We can now easily compare these numbers with each other in a meaningful way and uncover some interesting tendencies in the data. To highlight these tendencies, we computed the average relative normalized frequency of (un)attended *this* across all disciplines ($n_{average}$) and their standard deviations (*s*).

Table 7. Relative normalized frequencies (n) of (un)attended *this* across disciplines in MICUSP_Jan09

Discipline	Attended *this*	Unattended *this*
Biology	52	15↓
Civil & Environmental Engineering	71↑	17
Classical Studies	66	25
Economics	66	17
Education	52	20
English	52	16
Industrial & Operations Engineering	56	20
Linguistics	57	23
Mechanical Engineering	55	17
Natural Resources & Environment	47↓	20
Nursing	55	23
Philosophy	67↑	41↑
Physics	53	22
Political Science	46↓	18
Psychology	53	22
Sociology	56	21
$n_{average}$	56	21
s $n_{average}$	7	6
$n_{average}$ $+s$	63	27
$n_{average}$ $-s$	49	15

In Table 7, we have highlighted any frequency that is higher than the average plus one standard deviation with an arrow pointing up to indicate that the value is considerably higher than the overall average. By analogy, an arrow pointing down indicates that the value is considerably lower than the overall average (namely lower than the average minus one standard deviation). The mean values and corresponding standard deviations are given at the bottom of Table 7.

Table 7 reveals a rather diverse picture. For instance, we can see that in Biology papers, there is a tendency to avoid unattended *this*, and similarly, in Civil and Environmental Engineering, students very strongly prefer attended noun phrases. While these results could reflect a need for precise reference in the hard sciences, we find a comparable average of attended *this* in Philosophy texts. Interestingly, the latter are characterized not only by a frequent use of attended *this*, but also unattended *this* structures. In Natural Resources and Environment and Political Science essays, on the

other hand, the average number of attended *this* structures is below average. Overall, the results suggest that the cross-disciplinary differences in the use of (un)attended *this* are minor and relatively unsystematic. Contrary to what we may have expected, they cannot be accounted for solely with reference to differences between soft and hard sciences, but seem to indicate very discipline-specific stylistic preferences.

4.3.2. Nouns attending *this*

Let us now return to the question which nouns most frequently attend the demonstrative determiner *this*, and let us see if we can uncover clearer discipline-specific tendencies here. First of all, for the 6,827 cases of attended *this*, how many different nouns do we find? And are there any discipline-specific differences with regard to the variety of nouns? Table 8 provides the answer to these questions. For every discipline, we see (from left to right) the number of cases of attended *this* (i.e., tokens); the number of different nouns or noun phrases (so-called types); and a standard corpus-linguistic measure, the so-called type/token ratio (TTR). The TTR can be interpreted as a measure of how flexible or fixed the students' vocabulary is (within the limits of the particular structure we are concerned with here): the higher the TTR, the more different nouns serve to attend *this*; the lower the TTR, the less variation we find. For ease of comparison, the disciplines in Table 8 are sorted in order of descending TTR.

Table 8 confirms that disciplines do exhibit considerable variation: while the Classical Studies subsection of MICUSP_Jan09 has a TTR of 70%, that of the Nursing subsection is nearly half as large (36.02%). As with the general distribution of attended and unattended *this*, however, it appears that the TTRs do not fall into any coherent groups: Physics, clearly a hard science, has the second highest TTR (65.67%), while the Engineering and Biology texts rank somewhere between the middle range and bottom range TTRs. Similarly, the departments that belong to the Humanities and Arts, including English, Linguistics, and Education, display varying TTRs.

Are these results in fact suggesting that disciplines as remote as Classical Studies and Physics are actually much more similar in terms of writing styles than we might have assumed? In order to ultimately answer this question, we would have to engage in a detailed functional analysis of the noun phrases in question and take a closer look at the actual preferred noun types and recurring phrases (using an n-gram- or cluster-approach as described in Section 3.4). Similar TTRs may indeed reflect quite different functions of attended *this* as mentioned in Section 4.1 above: one is to avoid ambiguity, effectively paraphrasing referents already established in the preceding text; another is to offer interpretive or evaluative stance, that is, expressing ideas at a meta-level above the preceding text. It stands to reason if, and to what extent, students in, say, Classical Studies and Physics employ attended *this* for the same reasons.

Table 8. Head noun tokens, types, and type/token ratio (TTR) attending *this* in MICUSP_Jan09 by discipline

Discipline	Noun tokens	Noun types	TTR (%)
Classical Studies	150	105	70.00
Physics	67	44	65.67
Natural Resources & Environment	315	185	58.73
English	935	466	49.84
Sociology	278	138	49.64
Political Science	684	331	48.39
Civil & Environmental Engineering	284	126	44.37
Mechanical Engineering	185	82	44.32
Psychology	678	296	43.66
Linguistics	484	201	41.53
Industrial & Operations Engineering	383	158	41.25
Economics	324	129	39.81
Biology	631	251	39.78
Education	423	168	39.72
Philosophy	473	185	39.11
Nursing	533	192	36.02
Totals	6,827	3,057	

While space does not permit a fully-fledged functional analysis of the noun phrases across the disciplines, a frequency list of the most common nouns attending *this* across the disciplines provides valuable first insights. In analogy to Table 3 above, let us start with an overview of the top four nouns attending *this* in the 16 MICUSP_Jan09 disciplines; Table 9 provides these together with their absolute frequencies.

Table 9 largely confirms what Swales (2005) found in the Hyland corpus: with regard to the most frequent nouns, metadiscoursal and methodology-related nouns populate the top ranks, regardless of the specific discipline. Table 10 drives home the same point, adopting a slightly different angle: it displays the 25 most widely dispersed head nouns together with their absolute total frequencies.

Table 9. Most frequent nouns attending *this* in MICUSP_Jan09 by academic discipline

Biology	experiment (46)	study (25)	species (18)	paper (14)
Civil & Environmental Engineering	task (22)	report (15)	activity (13)	study (11)
Classical Studies	discussion (8)	paper (6)	point (6)	respect (5)
Economics	paper (44)	analysis (24)	model (12)	relationship (11)
Education	lesson (46)	paper (18)	study (14)	activity (13)
English	point (31)	sense (24)	passage (21)	way (18)
Industrial & Operations Engineering	study (30)	project (29)	data (18)	analysis (11)
Linguistics	paper (39)	experiment (18)	study (14)	project (13)
Mechanical Engineering	report (13)	method (12)	model (10)	study (9)
Natural Resources & Environment	study (31)	paper (13)	stream (8)	case (7)
Nursing	study (54)	project (32)	tool (23)	paper (21)
Philosophy	idea (26)	argument (24)	case (23)	paper (22)
Physics	paper (8)	scheme (4)	time (4)	information (3)
Political Science	case (28)	paper (26)	study (17)	period (14)
Psychology	study (44)	paper (37)	time (17)	idea (12)
Sociology	study (23)	paper (17)	point (12)	process (12)

Table 10 confirms that the most prominent nouns across the disciplines are metadiscoursal and methodology-related. In combination, Tables 9 and 10 demonstrate that the most frequent nouns are shared among the disciplines, and that there are a number of nouns that are shared by the majority of these disciplines that occur quite frequently. This stands in accord with Swales's (2005) analysis of the Hyland corpus and therefore is further evidence that cross-disciplinary differences (in terms of noun selection) are negligible and that overall, our student writers have a firm grasp of academic writing conventions that are comparable to those attested in the Hyland corpus, at least when it comes to the selection of nouns that attend *this*.

Table 10. Top 25 most widely dispersed head nouns attending *this* in MICUSP_Jan09

Head noun attending *this*	**n**	**Number of disciplines (out of 16)**
Paper	298	16
Case	136	15
Process	63	15
Point	98	14
Way	88	14
Section	52	14
Information	44	14
Study	277	13
Time	82	13
Issue	49	13
Problem	46	13
Method	42	13
Phenomenon	28	13
Situation	24	13
Idea	81	12
Model	81	12
Analysis	75	12
Sense	51	12
Question	50	12
System	50	12
Work	33	12
Data	58	11
Approach	38	11
Relationship	35	11
Difference	28	11

4.4 Summary of the findings

In our case study of (un)attended *this*, we used a range of corpus-linguistic methods to highlight different aspects:

- on the basis of a sorted concordance, we were able to quantify the shares of attended and unattended *this* in the 16 disciplinary subsections of MICUSP_Jan09; a chi-square test revealed that Philosophy texts are markedly different from all other disciplines in their relative overuse of unattended *this* (see also Wulff, Römer & Swales (Forthcoming));
- by resorting to normalized relative frequencies derived from our initial counts, we were able to uncover further discipline-specific trends which cut across disciplinary groups;
- by calculating the type-token ratios for the head nouns accompanying *this* in the different disciplines, we found that disciplines vary considerably with regard to the diversity of head nouns employed;
- a frequency list of the most common head nouns, combined with a list of the nouns most widely dispersed across the disciplines, supported earlier work on academic

writing, emphasizing the overall similarities between the disciplines in their use of attended *this* as a cohesive device that links preceding and subsequent argumentation by providing metadiscoursal links and initiating methodology-related explanation.

5. Concluding remarks

By way of exploring a pre-release version of the *Michigan Corpus of Upper-level Student Papers* (MICUSP_Jan09), this article has discussed how corpus linguistics and corpus methods can contribute to writing research, in particular research on advanced student academic writing. We have provided a basic introduction to what we consider core techniques in corpus analysis. A central aim was to show how software tools for corpus access enable users to see things that would be hard (or impossible, even) to see if the texts in a corpus were accessed without the help of such tools. One major advantage of a corpus/software-based approach to texts over a manual (non-computer-based) approach is that a much larger amount of language data can be examined in a short period of time, and new aspects about language (in our case student academic writing) can be captured and described. As Sinclair (1991) rightfully states, "[t]he language looks different when you look at a lot of it at once" (p. 100).

The usefulness of corpus analysis was then further exemplified through a case study of attended and unattended *this* in MICUSP_Jan09.[4] In this case study we saw that the creation and sorting of a concordance and the retrieval of information on the textual distribution of a word are powerful analytic techniques that may highlight usage patterns across disciplinary subsets of our corpus. The case study also demonstrated that it is important to treat disciplines separately in comparative analyses since groupings (e.g. according to academic divisions or faculties) may blur inter-disciplinary differences. While our case study was limited to one specific phenomenon ([un]attended *this*) in one specific genre (academic writing) by a specifically defined population (students), we hope to have given readers a taste of the possibilities offered by corpus linguistic methodology. Corpus methods are powerful in that they reveal patterns in the data that would otherwise escape the naked eye. While corpora and corpus tools will not do all the work for the writing researcher, they will help her/him discover phenomena that are worth investigating and highlight the preferred usage patterns of these phenomena.

References

Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamins.

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of *make* in native and non-native student writing. *Applied Linguistics*, *22*(2), 173-194.

Anthony, L. (2006). Developing a freeware, multiplatform corpus analysis toolkit for the technical writing classroom. *IEEE Transactions on Professional Communication*, *49*(3), 275-286.

Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.

Barlow, M. (2004). Software for corpus access and analysis. In J. M. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 205-221). Amsterdam: John Benjamins.

Barnbrook, G. (1996). *Language and computers*. Edinburgh: Edinburgh University Press.

Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book*. Boston: Heinle & Heinle.

Ellis, N. C., & Larsen-Freeman, D. (2009). Constructing a second language: Analyses and computational simulationsof the emergence of linguistic constructions from usage. *Language Learning*, *59*(1), 93-128.

Geisler, C., Kaufer, D. S., & Steinberg, E. R. (1985). The unattended anaphoric "this". *Written Communication*, 2, 129-155.

Goldberg, A. E. (2006). *Constructions at work. The nature of generalization in language*. Oxford: Oxford University Press.

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics, vol. 3: Speech acts* (pp. 41-58). New York: Academic Press.

Hoey, M. (2005). *Lexical priming. A new theory of words and language*. London: Routledge.

Huckin, T. N., & Olsen, L. A. (1991). *Technical writing and professional communication for non-native speakers of English*. 2nd edition. New York: McGraw-Hill.

Hunston, S., & Francis, G. (2000). *Pattern grammar. A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

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

Römer, U. (2009a). English in academia: Does nativeness matter? *Anglistik: International Journal of English Studies*, 20(2), 89-100.

Römer, U. (2009b). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, *7*, 141-163.

Schönefeld, D. (1999). Corpus Linguistics and Cognitivism. *International Journal of Corpus Linguistics*, *4*(1), 131-171.

Semino, E., & Short, M. (2004). *Corpus stylistics: speech, writing, and thought representation in a corpus of English writing*. London: Routledge.

Sinclair, J. M. (1991). *Corpus concordance collocation*. Oxford: Oxford University Press.

Swales, J. M. (2005). Attended and unattended "this" in academic writing: A long and unfinished story. *ESP Malaysia*, *11*, 1-15.

Swales, J. M., & Feak C. B. (2004). *English in today's research world: A writing guide*. Ann Arbor, MI: University of Michigan Press.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* Amsterdam: John Benjamins.

Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Cambridge, MA/London: Harvard University Press.

Wulff, S., & Römer, U. (2009). Becoming a proficient academic writer: Shifting lexical preferences in the use of the progressive. *Corpora*, *4*(2), 115-133.

Wulff, S., Römer, U., & Swales, J. M. (Forthcoming). Attended/unattended *this* in academic student writing: quantitative and qualitative perspectives. In E. Csomay. (Ed.), *Contemporary perspectives on discourse and corpora: New registers, analyses, texts, and tools* (Special issue of *Corpus Linguistics and Linguistic Theory*). Berlin: Mouton de Gruyter.

Acknowledgements

The authors would like to thank John Swales, Erin Friess and Ryan K. Boettger for helpful comments on earlier versions of this paper. Thanks are also due to Matthew Brook O'Donnell for his help with the MICUSP_Jan09 file preparation.

Notes

1. The URL for the MICUSP search and browse interface, MICUSP Simple, is http://search-micusp.elicorpora.info/.
2. The URL for Laurence Anthony's homepage is http://www.antlab.sci.waseda.ac.jp/. Different versions of *AntConc* can be downloaded from the 'software' section or directly from the *AntConc* homepage at http://www.antlab.sci.waseda.ac.jp/antconc_index.html.
3. Another factor that Swales suggests to play a role is the question whether the noun phrase is established in the preceding discourse, as in (3b), or not, as in (3c). While we do not consider this factor in the present case study, see Wulff, Römer, & Swales (Forthcoming).
4. Readers who are interested in further case studies based on subsets of MICUSP are referred to Römer (2009a), Römer (2009b), and Wulff & Römer (2009). These studies focus on phraseological items (n-grams and phrase-frames), introductory *it* patterns, and progressives, respectively.