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)

Ute Römer[1] and Matthew Brook O'Donnell[1]

Abstract

In this paper, we provide a detailed account of the steps that were central to designing and compiling the Michigan Corpus of Upper-level Student Papers (MICUSP). MICUSP is a new collection of 829 papers (around 2.6 million words) written by University of Michigan students in their final undergraduate year or in their first three years of graduate education. The papers come from sixteen disciplines, ranging from Humanities and Arts to Physical Sciences, and represent a range of different text types. In this paper, we offer an overview of the design of MICUSP, the online submission process used to collect papers, and the text-type classification of the papers.

1. Introduction

The Michigan Corpus of Upper-level Student Papers (MICUSP) is a new corpus of proficient student academic writing samples compiled by a team of researchers and students at the English Language Institute of the University of Michigan, Ann Arbor. The corpus, the first of its kind in North America, enables corpus researchers, EAP teachers and language testers to investigate the written discourse of proficient, advanced-level native- and non-native-speaker student writers at a large American research university.[2] It also provides students with a wide selection of A-graded papers that may serve as models for their own academic writing.

Papers of different types, ranging from essays to lab reports, have been collected from a wide range of disciplines within four academic

[1] Department of Applied Linguistics and ESL, Georgia State University, 34 Peachtree Street, Suite 1200, Atlanta, GA 30303, USA.
Correspondence to: Ute Römer, *e-mail*: uroemer@gsu.edu

[2] For more information about the background of the MICUSP project, we refer the reader to Ädel and Römer (forthcoming).

Corpora 2011 Vol. 6 (2): 159–177
DOI: 10.3366/corp.2011.0011

www.eupjournals.com/cor

divisions (Humanities and Arts, Social Sciences, Biological and Health Sciences, and Physical Sciences). The papers included in MICUSP were written by students at 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 of student writing. Around 830 papers have been collected through an online submission database that captures the required metadata pertaining to the student's linguistic background and academic experience. We also carried out a systematic analysis of the papers to classify their text type (e.g., 'argumentative essay', 'report').

In this paper, we discuss the steps that were central in the compilation of MICUSP. We provide an overview of the composition of MICUSP, the online submission process used to collect papers, and the text type classification of the papers. In a separate article (O'Donnell and Römer, forthcoming) we discuss the markup and annotation of the corpus and its distribution through the free online search and browse interface, MICUSP Simple.[3]

2. MICUSP paper collection and corpus composition

2.1 Text collection process

As we noted above, papers for MICUSP were collected from a wide range of disciplines – sixteen altogether. The selection of disciplines was largely influenced by MICUSP team members' previous experience in compiling the Michigan Corpus of Academic Spoken English (MICASE).[4] We were able to use some existing contacts in departments who had been supportive of MICASE and, in selecting disciplines for MICUSP we aimed to match MICASE to a large extent. In this way, we hoped to enable researchers to carry out comparisons between academic speech and academic writing within selected disciplines. As in MICASE, we used the University of Michigan's Academic Division categories – Humanities and Arts, Social Sciences, Biological and Health Sciences, and Physical Sciences – and selected between three and five disciplines from each category. The selected disciplines in MICUSP are: Biology, Civil and Environmental Engineering, Economics, Education, English, History and Classical Studies, Industrial and Operations Engineering, Linguistics, Mechanical Engineering, Natural Resources and Environment, Nursing, Philosophy, Physics, Political Science, Psychology, and Sociology.

In order to advertise the project and solicit submissions from students, we used a number of different strategies, including the distribution

[3] See: http://search-micusp.elicorpora.info/
[4] See: http://micase.elicorpora.info/

of flyers in prominent locations on campus, setting up information desks on campus during open days, and sending out mass e-mails to student mailing lists.[5] Of these strategies, we found that mass e-mailing worked best by far. As an incentive to submit and to compensate the student for their time and effort, we offered $10 gift certificates from a local book store.

When the project was launched in late 2004, the paper submission typically involved the following steps: in response to a MICUSP solicitation e-mail, a student e-mails his or her paper(s) to the MICUSP project manager; the project manager replies to the student's e-mail and suggests a date and time for an appointment; the student meets the project manager in her office to sign the consent form and complete a questionnaire with information about himself or herself and his or her paper; and the student receives a $10 Border's gift card. This was a very time-consuming procedure and paper collection was slow. In late 2006, the submission process was made much more efficient by setting up a web-based submission system. Students found the link to the online submission page in our solicitation e-mails. They simply had to click the link, log in with their University of Michigan user name, and follow a couple of steps outlined on the website (see screenshots in Figures 1 and 2). The uploading of papers, the filling in of the contributor questionnaire, and the signing of the consent form were all done electronically through this system, and meetings with the project manager were no longer required. The online submission system was implemented using the Ruby on Rails web application framework connected to a MySQL database. The key tables recorded data concerning the student (Contributor table) and information about the paper itself (Paper table). Once the student had submitted all of his or her information and the uploaded papers had been saved in our database, one of the MICUSP team members informed the student (by e-mail) about where to pick up the gift card.

Before submitting a paper for MICUSP, the students had to verify that they had handed in their paper with an instructor at the University of Michigan, that they had received an A or A– grade for their paper,[6] that they were a first- to third-year graduate student or a fourth year (senior) undergraduate when they submitted their paper for grading, and that they wrote the paper for a course in one of the sixteen departments listed above. The paper(s) had to be between 500 and 10,000 words long. Ten-thousand words was also the maximum number of words we accepted from each student (it was possible for a student to submit more than one paper, so long as he or she did not exceed the limit of 10,000 words).

Figures 1 and 2 illustrate the kind of information we collected about the students and their texts. The student completed the forms mainly

[5] To some extent the paper collection procedure described here is similar to that followed in the BAWE project (see Alsop and Nesi, 2009: 76–77).

[6] We asked the students to give the names of the instructor of the course and said that we may contact them in case we need to verify their grade.

Figure 1: Screenshot of the MICUSP paper submission interface (student information)

by ticking boxes or clicking on radio buttons, and by selecting options from drop-down menus. They first selected their gender and age range, and specified their 'department of concentration', (i.e., the department they were primarily affiliated with). They then answered a few questions about their language background and entered information about the text they uploaded. The students also selected from four options (senior undergraduate, first/second/third year graduate) what their standing was when they handed in their paper, how much time they spent approximately researching and preparing the text (e.g., 1–3 days, 2–3 weeks, 1 full semester), and whether they received feedback from someone on the text (e.g., from their instructor, a writing tutor or a classmate). From a drop-down menu they selected a text category for their paper. Options given were 'response paper', 'literature review', 'term paper', 'case study', 'technical or lab report', 'research proposal' and 'other' (plus explanation in a comment box). In a final step, the students gave informed consent by signing an electronic form to confirm that they had read the terms, and that they understood the nature of the study and agreed to take part in it. They were also given an opportunity to ask questions.

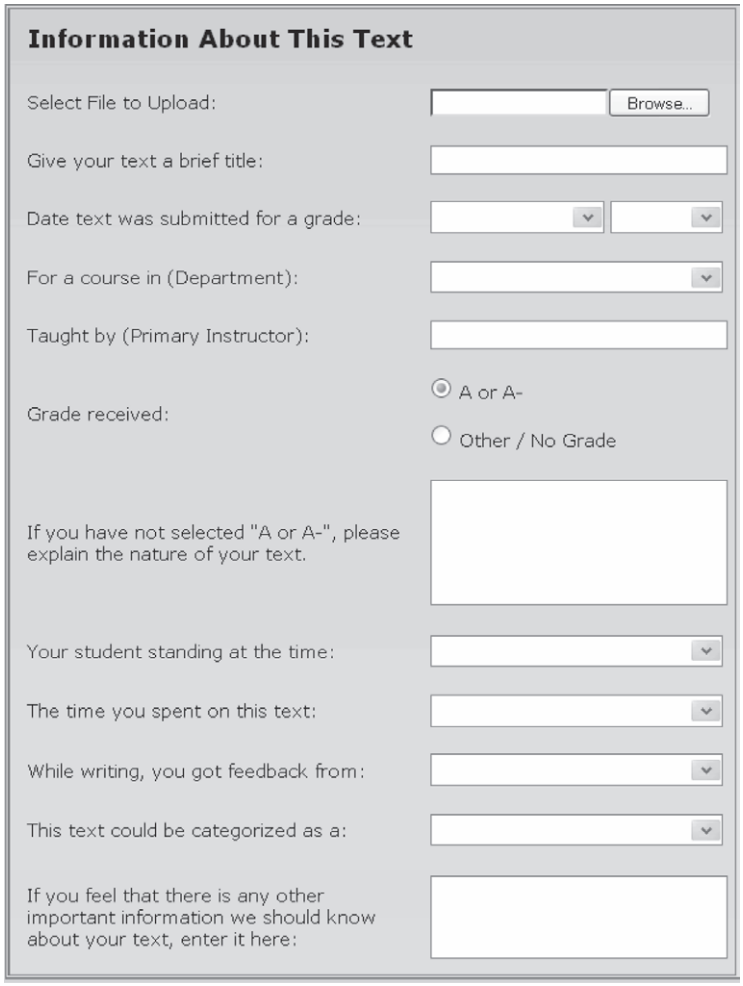


Figure 2: Screenshot of the MICUSP paper submission interface (paper information)

2.2 Corpus composition

MICUSP contains 829 student papers from sixteen disciplines and four student levels, making up over 2.6 million words. Tables 1 to 4 provide an overview of the distribution of MICUSP papers across disciplines, the level of the student within the institution, whether the student was a native or non-native speaker of English, and their gender.

As Table 1 shows, the numbers of papers from the individual disciplines range from 21 (for Physics) to 104 (for Psychology). While it was harder to get students in the Physical Sciences, and Biological and Health Sciences to submit papers, we managed to collect relatively large

Academic division	Discipline	Papers	Tokens
Humanities and Arts	English	98	268,733
	History and Classical Studies	40	182,629
	Linguistics	41	155,047
	Philosophy	44	128,028
		Σ 223	Σ 734,437
Social Sciences	Economics	25	78,070
	Education	46	150,282
	Political Science	62	210,783
	Psychology	104	323,326
	Sociology	72	215,793
		Σ 309	Σ 978,254
Biological and Health Sciences	Biology	67	176,124
	Natural Resources and Environment	62	176,653
	Nursing	42	158,773
		Σ 171	Σ 511,550
Physical Sciences	Civil and Environmental Engineering	31	98,918
	Industrial and Operations	42	124,973
	Engineering	32	123,335
	Mechanical Engineering	21	45,062
	Physics	Σ 126	Σ 392,288
All divisions and disciplines		Σ 829	Σ 2,616,529

Table 1: Distribution of papers across academic divisions and disciplines

sets of papers from disciplines in the Social Sciences and the Humanities and Arts, especially from Psychology, English and Sociology. Also, papers tend to be longer (in terms of word counts, not necessarily in terms of length by number of pages) in the Humanities and Social Sciences. The overall word count is highest for the Social Sciences division (978,254), followed by the Humanities and Arts (734,437), and Biological and Health Sciences (511,550), and it is lowest for Physical Sciences (392,288). At the outset, we had intended to collect the same number of papers from each discipline. One reason for the smaller number of papers from the 'hard sciences' could be that, even though the call for submissions only asked for 'any kind of already-written texts', some students in the hard sciences may not have thought that the types of assignments they produced (containing lots of graphs, figures, formulas, *etc*.) were what we were looking for. Besides, student numbers, and, in particular, the numbers of graduate students, vary greatly from discipline to discipline, so a small number of submissions from Mechanical Engineering may simply reflect low enrolment figures. Another possible explanation could be that students in some disciplines just write more than students in other disciplines. This is also what Nesi *et al*. (2004: 443) suspect to be the reason for the disciplinary imbalance in the British Academic Written English corpus (BAWE), which they consider 'a reflection

Student level		Papers	Tokens
Graduate	*third year*	77	359,092
	second year	117	446,336
	first year	203	747,747
	Total:	*397*	*1598,175*
Final year undergraduate		432	1,063,354
All levels		Σ 829	Σ 2,616,529

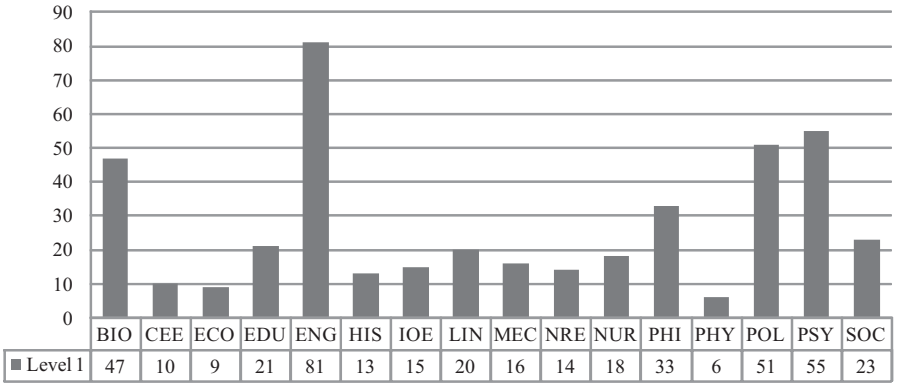
Table 2: Distribution of papers across student levels

of the fact that students in the humanities and social sciences produce more written work than science students.'[7]

If we divide all MICUSP papers by the level of the student when they submitted their papers, the distribution is somewhat more even than for the disciplines. We have an almost even division into 432 (52.1 percent) undergraduate and 397 (47.9 percent) graduate student papers. Note that graduate student writing accounts for almost 60 percent of the word tokens. Among the three graduate levels, figures decrease from first to third year, with 203 papers (24.5 percent) written by first-, 117 (14.1 percent) written by second-, and 77 (9.3 percent) written by third-year graduate students (see Table 2). Again, this uneven distribution of papers across student levels mirrors, at least to some extent, the different size of student populations at these levels. It could also indicate that upper-level graduate students either submit fewer texts for grades or are less willing to share them, or that the $10 gift card we offered was less of an incentive for them than for their lower-level peers.

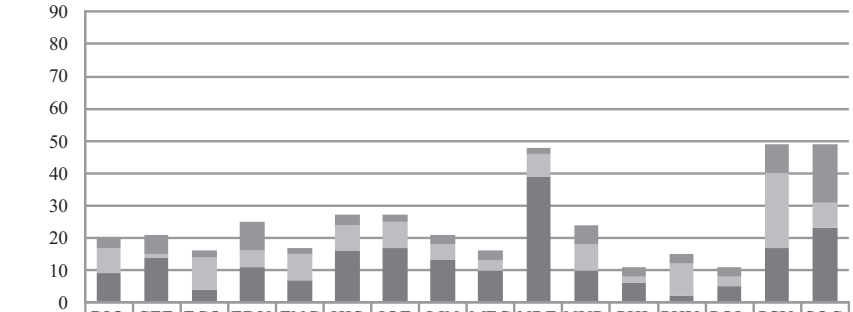
Each of the four levels shows a somewhat different distribution of papers across disciplines, as the charts shown under Figures 3 and 4 illustrate. Compared with the overall distribution (see Table 1), there are relatively more papers from Biology, English, Philosophy, and Political Science, and fewer from Civil and Environmental Engineering, Industrial and Operations Engineering, Natural Resources and Environment, and Sociology in the set of final-year undergraduate submissions (see Figure 3). Among the graduate student submissions, there are more, overall, from Civil and Environmental Engineering, History and Classical Studies, Industrial and Operations Engineering, Natural Resources and Environment, and Sociology, and fewer from Biology, English, Philosophy, and Political Science, again

[7] They subsequently modified their collection procedure in an attempt to reduce disciplinary imbalance. In the initial phases, it was decided to allow MICUSP to reflect to some degree the disparities in the amount of writing that goes on in different disciplines and the size of those disciplines, as this is instructive in itself. An attempt could be made in the future to balance the disciplines by either collecting more papers from low frequency disciplines or by extracting a random balanced sample as a subset.



	BIO	CEE	ECO	EDU	ENG	HIS	IOE	LIN	MEC	NRE	NUR	PHI	PHY	POL	PSY	SOC
Level 1	47	10	9	21	81	13	15	20	16	14	18	33	6	51	55	23

Figure 3: Distribution of final year undergraduate papers across disciplines



	BIO	CEE	ECO	EDU	ENG	HIS	IOE	LIN	MEC	NRE	NUR	PHI	PHY	POL	PSY	SOC
Level 4	3	6	2	9	2	3	2	3	3	2	6	3	3	3	9	18
Level 3	8	1	10	5	8	8	8	5	3	7	8	2	10	3	23	8
Level 2	9	14	4	11	7	16	17	13	10	39	10	6	2	5	17	23

Figure 4: Distribution of first through third year graduate papers across disciplines

compared to the distribution for MICUSP as a whole (see Figure 4, where the same scale is used as under Figure 3, to allow for accurate comparisons to be made). If we consider the three graduate student levels separately, we see that among the submissions from History and Classical Studies, Industrial and Operations Engineering, Natural Resources and Environment (in particular), and Sociology, there are comparatively high numbers of papers that were written by first-year graduate students. Second-year graduate student papers are somewhat over-represented in the set of Psychology papers and are extremely rare in the sets of Civil and Environmental Engineering, and Philosophy papers. The highest share of third-year graduate student papers can be found among the Sociology submissions.

Nativeness status	Papers	Tokens
Native speaker	681	2,173,030
Non-native speaker	148	443,499
All	Σ 829	Σ 2,616,529

Table 3: Distribution of papers across nativeness groups

Gender	Papers	Tokens
Male	314	1,007,960
Female	515	1,608,569
All	Σ 829	Σ 2,616,529

Table 4: Distribution of papers across student gender

As Table 3 shows, about 82 percent of MICUSP papers (681) were written by students who gave English as their native language (or as one of two native languages, in the rare case of bilingual contributors). At 148 (17.9 percent), the set of papers produced by non-native speakers of English is considerably smaller. This number should, however, still be large enough to provide a useful basis for comparative analyses of proficient native and non-native student writing. The figures under Table 4 indicate that, on the whole, female students were more prepared to submit papers than male students. Of the 829 papers in MICUSP, 515 (62.1 percent) were written by women and 314 (37.9 percent) by men. The higher number of female-authored submissions probably relates to the larger numbers of papers from Humanities and Social Sciences disciplines where we find, on average, more women enrolled than men.

3. MICUSP paper classification

In the process of submitting papers for MICUSP, students were asked to specify for each text submitted how it might be categorised. They were able to select a paper category from a drop-down menu or select 'other' and provide a label themselves. It appears that the majority of students had labelled their texts 'term paper' (one of the categories in the list), and a random check of papers from the database indicated that the students' self-categorisation of texts was, overall, not very reliable (see also Alsop and Nesi, 2009: 76). We therefore decided to examine the entire collection of papers ourselves

and develop a system of paper classification based on what we found in our dataset.

The main goal of the MICUSP paper classification was to provide a further option of subdividing the set of 829 papers in a meaningful way (in addition to subdivisions according to discipline, level, or other speaker characteristics) and thus enable corpus users to search only in, or browse for, papers of a particular type. Our paper classification system was developed through a series of interlocking steps that were carried out by a group of scholars and graduate students in the fields of corpus linguistics, genre analysis, EAP pedagogy, and language testing.[8]

3.1 Determination of paper categories

In our first step we reviewed a range of classifications and definitions of academic text-types used in the EAP literature (e.g., Johns, 2002; Paltridge, 2004; and Swales and Feak, 2004) and provided by writing clinics and student writing support divisions of a number of US universities with strong programmes in academic writing (including Duke University, the University of Michigan and the University of North Carolina). We also considered the categorisation system used in the BAWE project.[9] From what we were able to find, it seems that there appears to be little consensus on what counts as central types of student writing or on how listed categories (e.g., 'report' or 'argumentative essay') ought to be defined. Furthermore, we found that no existing set of categories was able to reflect what was present in our multi-disciplinary/cross-level dataset. Definitions were often restricted to particular disciplines, while our aim was to develop a set of 'maximally inclusive' categories. For example, we wanted to capture in our 'report' category reports from a range of different disciplines, not just lab reports in Biology or technical reports in Mechanical Engineering. So we decided to develop a set of paper categories and definitions that fully mirrored the composition of the MICUSP dataset, rather than adopt any of the existing text-type sets.

Our development of a classification system was essentially data driven. Random sets of MICUSP papers were pulled from the database and classified independently by a group of linguists, EAP teachers and graduate students in English and Education. In a first group meeting, all individual classifications were discussed, and we devised an initial set of paper categories and definitions. Based on more evidence from MICUSP (further sets of randomly selected papers), the categories and definitions were

[8] The authors would like to thank John Swales, the members of his Winter 2009 independent study group (Laura Aull, Moisés Escudero, Tim Green and Zak Lancaster), and a group of ELI instructors and assessment specialists (Pamela Bogart, Chris Feak, Mindy Matice, India Plough, Sue Reinhart and Theresa Rohlck) for their valuable discussions in the paper classification process and for their help with multiple revisions of the paper-type definitions.
[9] See: www.coventry.ac.uk/bawe

revised several times in group discussions. In putting together the definitions, we listed what we considered to be the core features of each paper category, independent of the disciplines that the papers were written in. In this way, we developed a set of inclusive categories that capture some of the core types of student academic writing across different fields of study. The result was a list of seven paper categories: argumentative essay, creative writing, critique/evaluation, proposal, report, research paper and response paper. Our definitions and text type examples for each of these categories are given under Table 5. The definitions consist of a description of each category's rhetorical purpose and a set of defining features. The classification of MICUSP papers described under Section 4.3 is based solely on these definitions. This means that a paper is only coded as a 'response paper' if it is a response paper according to our definition – not according to other definitions found in the literature or the coder's personal understanding of what a response paper is.

3.2 Supplementary textual features for EAP teachers

It became clear from our conversations with EAP instructors at the University of Michigan English Language Institute that, very often, EAP teaching focusses on a particular section of a paper or discourse strategy – for example, paper abstracts or how to refer to tables or graphs. The instructor is then looking for sample papers which contain the section type or realisations of the discourse strategy in question. So, in addition to classifying texts into paper categories, we decided also to label them for a number of textual features that are important in teaching academic writing so that it would be easier to identify MICUSP papers that are relevant for specific teaching purposes.

We invited instructors to suggest textual features or sections they often focus on in their EAP/ESL classes. From the list of suggestions we then selected eight features that we regarded as codeable by our student research assistants. The eight selected textual features are: abstract, definitions, discussion of results, literature review, methodology section, problem–solution pattern, reference to sources, and tables, graphs or figures.

3.3 Paper classification procedure

The classification of MICUSP papers and the coding of textual features was done by means of an online user interface that was designed specifically for the purpose of the MICUSP paper classification (see screenshot under Figure 5). Our coders, mainly University of Michigan graduate students (with specialisations in Information, English, and Education), logged on to the website with a unique username and password, and were randomly assigned

Paper category	Rhetorical purpose	Features	Examples
Argumentative essay	demonstrates ability to construct a coherent argument and support it with evidence /examples	- paper is thesis driven - author's thesis is supported by pieces of evidence from an outside source - may generate a new idea/argument in the field	argumentative essay, persuasive essay, literary analysis essay
Creative writing	by definition, the texts in this category do not adhere to any particular rhetorical purpose or structure		narrative writing, poetry, drama scripts
Critique / evaluation	presents a positive and/or negative assessment of an outside source/project/ text	- text is driven by an in-depth assessment of a product/policy/ procedure/text (although often interwoven with a description or observation of the product/policy/ procedure/text) - gauges the effectiveness, validity, or usefulness of something - recommendations for improvement may be offered	evaluation of business practices, problem–solution, literary critique, operations report
Proposal	puts forth a research question, a theory, or a model that the author feels should be explored in order to further the understanding of a given topic	- formulates a research question or model, or proposes a potential study - usually does not collect or synthesise new data, but may include projected results; any collected data will be to support the proposal - justifies the need for data collection or data verification - critiques relevant literature and/or prior studies	research proposal, numeric model proposal, effective business/ management design
Report	describes the state or gives an account of a problem/issue/ text, or describes the carrying out of a procedure (demonstrates the ability to gather data and summarise)	- most space is devoted to description, rather than critical assessment - not driven by an original thesis or research question - author's opinion/evaluation may be present, but is not foregrounded and does not appear to drive the text	lab report, literature review, article review, annotated bibliography, compare / contrast paper

Table 5: MICUSP paper categories and definitions for paper classification

Paper category	Rhetorical purpose	Features	Examples
Research paper	presents original research in a field	- entire text serves to answer a clearly stated research question - contains original data, or compiles existing data for the purpose of providing a new interpretation - structured into predictable sections (usually with subheadings) - includes most of the following: abstract, literature review, methods, results, discussion, conclusion	research paper, replication study
Response paper	short piece of writing responding to a given prompt or question, although prompt may not be explicit in text	- short in length (typically 1–2 pages) - style tends to be informal (e.g., expressions of emotional response; frequent references to mental processes, such as 'I was confused', 'I was surprised') - may lack a formal introduction/ 'jumps right in' to content of paper, because author assumes reader's familiarity with the given topic (shared knowledge or in-group knowledge) - text provokes new questions for the author that may not be thoroughly answered	solution to a homework problem, personal response to a text

Table 5: (*continued*) MICUSP paper categories and definitions for paper classification

papers to classify. The papers were anonymised versions of MICUSP papers in the form in which they were originally submitted (with the original layout and structure maintained). Filenames had been replaced with random numbers in order to prevent the coder from getting clues from the discipline or student level encoded in the filename and, for example, being tempted to label every Mechanical Engineering paper 'report', knowing that the lab report is a common type of text in this particular discipline. The coder then scanned the paper he or she was assigned and read as much of it as was necessary to determine its paper category and to decide on the presence/absence of each of the eight textual features listed in the previous section. Paper categories were selected from a drop-down menu (see Figure 5). For the presence of textual features, a series of tick boxes were checked or left unchecked.

Each paper was classified independently by two coders. Two additional coders performed random checks. The coding procedure was double-blind, meaning that no coder was entitled to see the other coder's

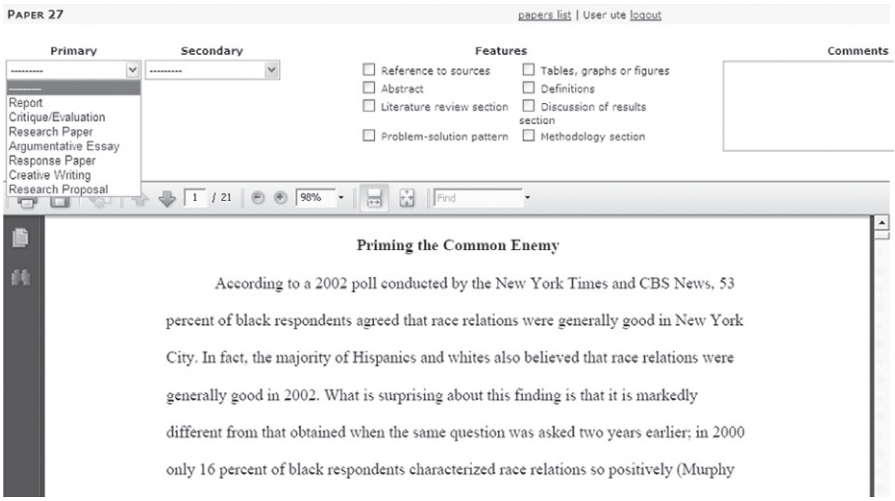


Figure 5: Screenshot of online interface designed for paper classification

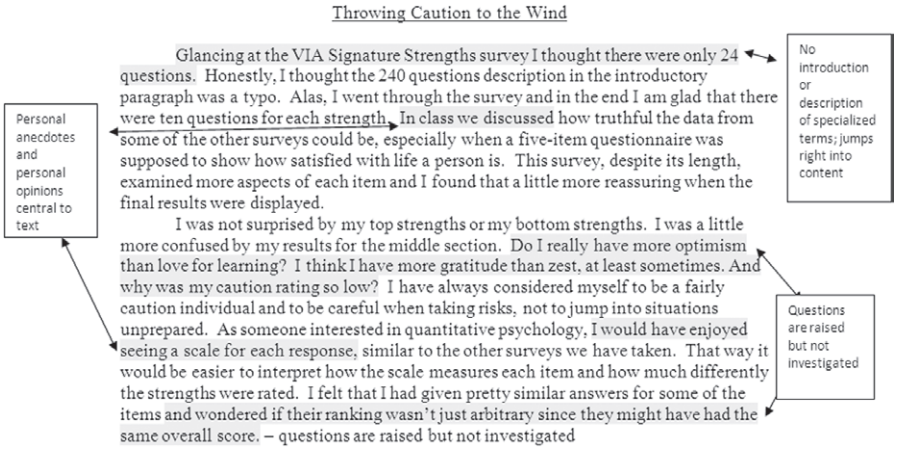


Figure 6: Example of a response paper with a few characteristic features highlighted

choices on paper categories or textual features. Papers received an obligatory primary and an optional secondary paper category label, based on the definitions given in Table 5. Marking a primary classification meant that the coder thought the paper fulfilled the rhetorical purpose of the selected category and that the paper contained most or all of the features listed under that category. Figure 6 illustrates how certain aspects of a text (here a response paper) helped a coder to make a decision on which category to select. Inter-rater reliability was determined on a weekly basis and ranged

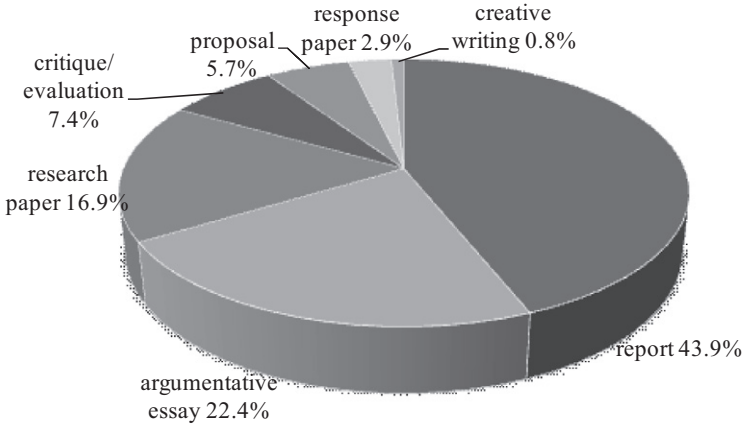


Figure 7: Distribution of MICUSP papers across categories

from 57.9 percent to 88.4 percent (with an average value of 74.6 percent and a standard deviation of 11.7), which means that, overall, coders agreed on the primary paper category in 74.6 percent of cases. Where there was no agreement between coders one and two, the paper was assigned to and classified by a third coder. If there was still no agreement on the primary paper-type at this stage, the paper was looked at and discussed by all coders in a group meeting. Such cases were rare and it was always possible to arrive at a consensus, given that decisions could be based on a carefully developed set of paper-type definitions.

3.4 Distribution of papers across categories and textual features

Figure 7 provides an overview of the distribution of the 829 MICUSP papers across paper types. As the figure shows, the most common paper category (with 43.9 percent or 364 texts) is the report. Another dominant paper type is the argumentative essay (22.4 percent, 186 texts), followed by research paper (16.9 percent, 140 texts) and critique/evaluation (7.4 percent, 61 texts). The remaining three categories (proposal, response paper and creative writing) are comparatively rare in our dataset. If we now split up all the papers that were classified as reports by discipline (see Figure 8), we see that roughly half of the papers in this category come from only five different disciplines: Psychology (fifty-three papers), Natural Resources and Environment (thirty-seven papers), Biology (thirty-one papers), Political Science (twenty-nine papers) and Sociology (twenty-eight papers). Nursing and Education also contribute larger numbers of texts to this category. This, however, does not necessarily mean that the report is the dominant student paper-type in all of these disciplines or that it is avoided by writers from

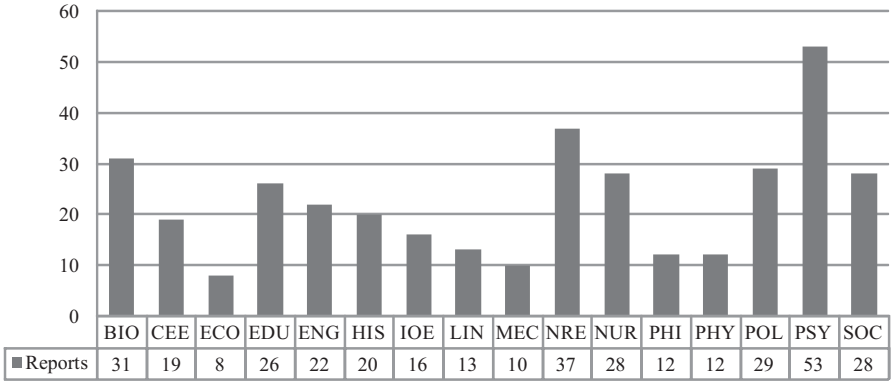


Figure 8: Distribution of reports across disciplines

the remaining disciplines. If we set the numbers given under Figure 8 in relation to those under Table 1 (i.e., the overall numbers of papers in each discipline), Psychology, Natural Resources and Environment, Biology, and Political Science still stand out by containing above-average shares of reports but Sociology does not (with 38.9 percent, or twenty-eight out of seventy-one papers, the percentage is below average). Besides, Civil and Environmental Engineering (61.3 percent reports), History and Classical Studies (50 percent reports) and Physics (57.1 percent reports) show a clear preference for this type of paper although these disciplines do not top the list, based on absolute numbers of reports.

MICUSP papers can also be split up according to which of the supplementary textual features given under Section 3.2 they contain. Table 6 provides the overall shares, and shares by student level, of each of the eight textual features we identified as being of interest to EAP/ESL teachers. MICUSP papers containing an abstract account for 12.1 percent of cases. Definitions were found in 10.7 percent of the papers. A discussion of results section was found in 24.1 percent of all papers, and 12.7 percent have a literature review. A methodology section is present in 21 percent of the papers, while the problem–solution pattern is rather rare (with only 10.4 percent of MICUSP papers containing one). References to sources were found in 75 percent of the papers, and 29.7 percent contain either tables, graphs or figures. As Table 6 shows, there are some interesting differences with respect to the presence of the eight selected textual features across the four student levels. We observe a steady increase in relative frequencies from final year undergraduate to third year graduate student writing with respect to the presence of definitions (8.6 percent to 16.9 percent), literature review sections (8.8 percent to 26 percent), and references to sources (24.3 percent to 35.2 percent). Of these three textual features, only the literature review showed significant cross-level differences (according to the chi-square test). We still think that the observed differences in the use of references to

	Final year undergrad.	1st year graduate	2nd year graduate	3rd year graduate	MICUSP_all
Abstract	10.9	14.8	8.5	16.9	12.1
Definitions	8.6	11.3	13.7	16.9	10.7
Discussion of results	24.1	25.6	22.2	23.4	24.1
Literature review	8.8	12.8	17.9	26.0	12.7
Methodology section	20.1	22.7	20.5	22.1	21.0
Problem–solution pattern	7.2	13.3	15.4	13.0	10.4
Reference to sources	69.7	76.8	84.6	85.7	75.0
Tables, graphs or figures	24.3	37.4	32.5	35.1	29.7

Table 6: Shares of MICUSP papers across levels containing supplementary textual features (figures shown are percent)

sources and definitions merit a closer examination. Also significant according to chi-square are the observed differences between the level-specific counts for the presence of the problem–solution pattern and tables, graphs or figures. The problem–solution pattern is relatively frequent in second-year graduate papers and less common in final-year graduate submissions. Tables, graphs or figures are used in a significantly higher percentage of graduate student papers (especially first year) than undergraduate student papers.

4. Conclusion

Compiling a corpus as a public domain resource for different groups of users (including linguistic researchers, EAP teachers and students) is not a trivial matter. It involves extensive resources, careful and detailed planning, and a team of people with the right mix of technical skills and linguistic knowledge. In this paper, we have provided an account of the design and compilation of MICUSP, a corpus of upper-level student papers from different academic disciplines. We have discussed important issues related to MICUSP text solicitation and collection, the composition of the corpus, and the classification of the papers according to text types. We have presented each of these issues in sufficient detail for future MICUSP users to learn enough about what is in the corpus and what information can be retrieved from it.

The corpus allows a range of investigations into the nature of student writing across academic fields and levels. A number of internal research

projects based on MICUSP have already been carried out or are under way, on topics including the use of (un)attended *this* (Römer and Wulff, 2010; and Wulff *et al*., forthcoming), scare quotes (Aull and Barcy, 2010; and Avanessian and Swales, 2010), attribution (Ädel and Römer, forthcoming), introductory *it* patterns (Römer, 2009b), progressives (Wulff and Römer, 2009), and phraseological items (Ädel and Römer, submitted; O'Donnell and Römer, submitted; and Römer, 2009a). On the pedagogical side, MICUSP is already being used by EAP instructors and their students in academic writing classes at the University of Michigan. We hope that a larger number of students, teachers and researchers will soon start to explore the corpus and discover exciting aspects of student academic writing.

Acknowledgements

MICUSP and MICUSP Simple have been very much a team effort. The authors would like to acknowledge the support of a number of people who were involved in the project at various stages (in alphabetical order): Annelie Ädel, Derek Blancey, Kate Boyd, Yung-Hui Chien, Gregory Garretson, Geoffrey Ho, Lucas Jarmin, Miranda Kozman, Emily Lin, Kelly Lockman, Nasy Pfanner, Jesse Sielaff, Rita Simpson-Vlach, John Swales, Madison Stuart, Edwin Teng and Beilei Zhang. We are also grateful for the support of instructors and testing specialists at the University of Michigan English Language Institute and members of the University of Michigan Corpus Analysis Group.

References

Ädel, A. and U. Römer. forthcoming. 'Research on advanced student writing across disciplines and student levels: introducing the Michigan Corpus of Upper-level Student Papers', International Journal of Corpus Linguistics.

Alsop, S. and H. Nesi. 2009. 'Issues in the development of the British Academic Written English (BAWE) corpus', Corpora 4 (1), pp. 71–83.

Aull, L. and K. Barcy. 2010. 'The "good", the "bad", and the snarky: native and non-native student use of scare quotes in upper level academic writing'. Paper presented at the 6th Conference on Intercultural Rhetoric and Discourse. June 2010. Georgia State University, Atlanta, GA.

Avanessian, N. and J.M. Swales. 2010. 'Scare-quotes in MICUSP: some preliminary observations'. MICUSP Kibbitzer. Accessed 30 July 2010, at: http://micusp.elicorpora.info/micusp-kibbitzers/

Johns, A. (ed.). 2002. Genre in the Classroom: Multiple Perspectives. Mahwah, New Jersey: Lawrence Erlbaum.

Nesi, H., G. Sharpling and L. Ganobcsik-Williams. 2004. 'Student papers across the curriculum: designing and developing a corpus of British student writing', Computers and Composition 21, pp. 439–50.

O'Donnell, M.B. and U. 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 7 (1).

O'Donnell, M.B. and U. Römer. Submitted. 'Investigating the interaction between phraseological items and textual position'.

Paltridge, B. 2004. 'Academic writing', Language Teaching 37 (2), pp. 87–105.

Römer, U. 2009a. 'English in academia: does nativeness matter?', Anglistik: International Journal of English Studies 20 (2), pp. 89–100.

Römer, U. 2009b. 'The inseparability of lexis and grammar: corpus linguistic perspectives', Annual Review of Cognitive Linguistics 7, pp. 140–62.

Römer, U. and S. Wulff. 2010. 'Applying corpus methods to written academic texts: explorations of MICUSP', Journal of Writing Research 2 (2), pp. 99–127.

Swales, J.M. and C.B. Feak. 2004. Academic Writing for Graduate Students. Ann Arbor, Michigan: University of Michigan Press.

Wulff, S. and U. Römer. 2009. 'Becoming a proficient academic writer: shifting lexical preferences in the use of the progressive', Corpora 4 (2), pp. 115–33.

Wulff, S., U. Römer and J.M. Swales. Forthcoming. 'Attended/unattended *this* in academic student writing: quantitative and qualitative perspectives', Corpus Linguistics and Linguistic Theory.