Comparing Expert and Learner Mathematical Language: A Corpus Linguistics Approach

Lara Alcock[1], Matthew Inglis[1], Kristen Lew[2],
Pablo Mejia-Ramos[3], Paolo Rago[1] & Chris Sangwin[4]
[1]Loughborough University, [2]Arizona State University,
[3]Rutgers University, [4]University of Edinburgh

Corpus linguists attempt to understand language by statistically analyzing large collections of text, known as corpora. We describe the creation of three corpora designed to enable the study of expert and learner mathematical language. Our corpora were formed by collecting and processing three different genres of mathematical texts: mathematical research papers, undergraduate-level textbooks, and undergraduate dissertations. We pay particular attention to the method by which our corpora were created, and present a mechanism by which LaTeX source files can be easily converted to a form suitable for use with corpus analysis software packages. We then compare these three different types of mathematical texts by analyzing their word frequency distributions. We find that undergraduate students write in remarkably similar ways to textbook authors, but that research papers are substantially different. These differences are discussed.

Key words: corpus linguistics, mathematical language, proof

Understanding the nature of mathematical language is a goal for at least three research communities. Sociologists and philosophers have long been interested in the practices of intellectual communities, and mathematics, with its uniquely deductive mode of inquiry, has attracted particular attention (e.g., Larvor, 2016). Mathematicians increasingly recognize that novices need to learn not only the content of mathematics but also the practices of mathematicians. Transition-to-proof courses are therefore common, and a growing number of textbooks directly address logical norms of mathematical communication (e.g., Vivaldi, 2014). Mathematics educators at all levels aim to support learners in developing sophisticated modes of thinking: a commonly-stated goal is that learners should engage in authentic mathematical activity that involves reasoning, proving, and communicating their arguments with others in the classroom and in written work (e.g., Stylianides, 2007).

These communities – sociologists and philosophers, mathematics educators, and mathematicians – therefore share an interest in understanding the norms of mathematical practice and communication. To date, however, there are relatively few empirical studies of this practice, and those that exist indicate less homogeneity among mathematicians than is typically portrayed in introspective accounts (e.g., Inglis, Mejía-Ramos, Weber & Alcock, 2013). In particular, to our knowledge, there have been few large-scale attempts to study the authentic mathematical communication of research mathematicians, or to compare this to the communication of undergraduates.

One method of studying language is to use the techniques of corpus linguistics, a branch of linguistics that statistically interrogates large collections of naturally occurring text, known as corpora. Methods developed by corpus linguists can be used to investigate many different types of linguistic question, and have revealed important and surprising findings (e.g., McEnery & Hardie, 2011).

Our goal in this project was to compare three distinct types of mathematical written language: that used by mathematicians when writing research papers, that used by mathematicians when communicating with undergraduates for pedagogical purposes, and that used by undergraduates when writing for assessment. By collecting and processing naturally occurring mathematical texts of these three types we aimed to understand the similarities and differences of these three genres of mathematical language. A subsidiary goal was to compare all these versions of mathematical language with general (non-mathematical) written English. We first discuss the process involved in creating our three corpora.

Collecting the Texts

The first task for a researcher who wishes to create a corpus is to collect examples of the language that they wish to study. We adopted two largely pragmatic criteria:

1. We collected only text in LaTeX format to enable consistent processing (discussed below).
2. We collected only text that had been published non-commercially or, in the case of student projects, where the author agreed to assign us copyright.

Subject to these criteria, we collected texts for the three corpora in different ways.

To create the learner corpus we invited undergraduate students to submit their final-year projects or dissertations. Such dissertations are common in the UK (where students specialize to study three subjects at age 16 and to one or two when at university) and degree programs vary in their dissertation criteria. But students commonly have the opportunity to undertake an individual project – which might be expository or applied– accounting for approximately one sixth of their final-year credit. We invited submissions from such students via project coordinators at 15 universities, who sent on an email directing interested students to a Facebook page that explained how to process their dissertation LaTeX file to remove personal identifiers, and how to submit this along with a copyright transfer form. Each student who submitted received a £5 (approximately $7) Amazon voucher and was asked to encouraged their friends to submit too. By this method we collected 50 student dissertations, which contained a total of 419,965 words.

To create the pedagogic text corpus we located online undergraduate-level open textbooks using the Open Textbook Library, the College Open Textbooks site, and the American Institute of Mathematics Approved Textbook list. Topics included abstract algebra, analysis, linear algebra, complex analysis, and textbooks designed to support the transition to proof. If the textbooks were not available in LaTeX format we contacted the author and asked for permission to access their source files. This approach left us with the source files for 21 complete undergraduate textbooks, which contained a total of 1,518,932 words.

To create the expert corpora we first downloaded all papers that had been uploaded to the arXiv in the first four months of 2009. The arXiv is an online repository that is routinely used by research mathematicians to share their research articles. The majority of articles on the arXiv are available in LaTeX format, and can be bulk downloaded using a command line tool. We then sorted the articles using their primary subject classification (e.g., mathematics, physics, etc.) and further sorted them using their secondary subject classification (e.g., algebraic geometry, algebraic topology, etc.). This left us with a total of 6988 mathematics articles, containing 30,892,695 words.

Processing the Texts

Collecting mathematical language and converting it into a form that can be processed using the standard software packages used by corpus linguists presents a challenge. Unlike most texts, mathematical language contains numerous atypical characteristics, such as inline mathematical notation. Most mathematics is written using the LaTeX markup language, not plain text. Our first goal was therefore to create a method of converting LaTeX source code to plain text in a way that preserved the natural sentence structure of the language, but which removed non-linguistic features of the source code (the code for bold or italic text for instance).

An important question for the would-be creator of a mathematical corpus concerns how to deal with inline mathematical notation. For instance, a typical mathematical sentence might be "Let $f : X \rightarrow Y$ be a bijection." How would we want the "$f : X \rightarrow Y$" to appear in a plain text corpus? One approach would be to leave the LaTeX source code intact and to analyze the code as if it were natural language. The difficulty with adopting this option is that there are several different ways in which one could encode "$f : X \rightarrow Y$" in LaTeX. For instance, `$f:X\rightarrow Y$` and `\(f:X\rightarrow Y\)` produce identical output, and `$f\,:\,X\longrightarrow Y$` only differs stylistically. We therefore felt that this approach would be unhelpful for the majority of questions a researcher would wish to answer using a mathematical corpus (although our code does allow this approach as an option, as a researcher who wished to primarily focus on the semantic content of papers might wish to retain these markup codes).

A second option would be to delete all mathematical code entirely, and simply record the example above as "Let be a bijection". We rejected this option as it seemed not to preserve the logical structure of sentences, which would influence certain analyses (those that investigate the collocation of words, for instance). Instead we opted to replace all occurrences of inline mathematics with the string "inline_math" (although this decision can be altered by the researcher if desired). The scripts we used to convert LaTeX to analysis-ready plain text are freely available for the research community at:
`https://github.com/sangwinc/arXiv-text-extracter`

As our corpus of general written English we used the combined Lancaster-Oslo/Bergen corpus (commonly referred to as the LOB corpus; Johansson, 1986) and Brown corpora. The Brown University Corpus of Standard American English (commonly referred to as the Brown corpus; Francis & Kucera, 1961) is formed of 1 million words of American English from texts published in 1961. The LOB corpus consists of written British English created to mirror the structure of the Brown corpus (i.e. texts were taken from similar sources in similar proportions). Thus our combined Brown/LOB corpus consisted of 2 million words of British and American written English.

Analyzing the Corpora

Having created the corpora, our primary goal was to understand the extent to which they were similar: is it the case that the language used in mathematical textbooks, mathematics research papers, and undergraduates' final year projects is consistent? If not, where are the differences between these genres, and how can these differences be characterized?

Kilgarriff (2001) proposed a variety of measures that aimed to assess the similarity of different corpora. All his approaches relied upon the so-called 'bag of words' model of text construction. This model ignores the order in which words occur and instead focuses on understanding texts by assessing their distributions of word frequencies. The basic idea is that two texts are likely to be a similar genre, and focus on a similar topic, if they have broadly similar word frequency distributions. Of the measures he studied, Kilgarriff concluded that a chi-squared approach performed best. Suppose one wishes to calculate the similarity of corpora A and B. Kilgarriff proposed determining the most frequent *n* words in the supercorpus formed of $A \cup B$, and then calculating the test statistic for a chi-squared test of goodness of fit. Since these *n* words were selected to be the most frequent, and not sampled randomly from the population of words, it would be inappropriate to actually perform the chi-squared hypothesis test, but Kilgarriff reasoned that the test statistic would serve as a suitable measure of similarity (with lower values represent more similarity).

Unfortunately Kilgarriff's (2001) chi-squared method would not suffice for our purposes, as it requires that we are comparing corpora of the same size. We therefore modified his proposal as follows. We first determined the 100 most frequent words across our four corpora, where each corpus was weighted as representing 25% of the supercorpus (we needed to perform this weighting because our expert arXiv corpus was considerably bigger than the others). We did not include "inline_math" as a word for this analysis, as clearly it did not appear at all in the Brown/LOB corpus. We then calculated the proportion of each corpus consisting of each word. For instance, the word "the" represented 6.08% of the arXiv corpus, 6.72% of the textbook corpus, 6.62% of the learner corpus, and 6.69% of the Brown/LOB corpus.

For each pairwise combination of corpora, *A* and *B*, we then calculated

$$S_{AB} = \sum_{i=1}^{100} \left(\frac{(a_i - b_i)^2}{a_i} + \frac{(a_i - b_i)^2}{b_i} \right)$$

where a_i represents the proportion of corpus *A* formed of word *i*, and b_i represents the proportion of corpus *B* formed of word *i.* While this is not a true chi-squared value (which would be calculated with frequencies rather than proportions) it fulfils a similar role. Therefore if $S_{AB} < S_{AC}$, we can conclude that corpora A and B are more similar than corpora A and C. The S_{AB} values for each pairwise combination of our four corpora are given in Table 1, and plots of the frequencies of the top 100 words are shown in Figure 1 (so, a point at (*x*,*y*) in the bottom left graph indicates that the same word formed *x*% of the arXiv corpus and *y*% of the textbook corpus).

The results shown in Table 1 and Figure 1 paint a consistent picture. We found that the textbook and learner corpora had remarkably similar word frequency distributions, that the arXiv corpus formed of mathematical research papers was somewhat different, and that all three mathematical corpora were substantially different to the regular written English of the Brown/LOB corpus. Before exploring the differences between the arXiv and textbook corpora below, we first make some remarks on these findings.

	Textbook	Learner	Brown/LOB
ArXiv	0.105	0.097	1.215
Textbook		0.011	1.317
Learner			1.320

Table 1: The similarity measures, S_{AB}, for each pairwise combination of our four corpora.

Although our analysis was exploratory in the sense that we did not have strong hypotheses about the results in advance, we were somewhat surprised by these findings. We anticipated there would most likely be a gap between the language used by experts and novices. After all, mathematicians have typically had many years of enculturation into the discipline, whereas the undergraduates who provided the texts for our learner corpus had only had three or four years of university-level study. However, we found something quite different. Our learners seemed to produce very similar written language to that found in textbooks written by experts, at least in the sense that their word frequency distributions were close to identical. One hypothesis that might account for this similarity would be if the two corpora had similar balances of mathematical topics. For instance, two corpora focused on linear algebra might be expected to have similar word frequencies for “kernel”, “matrix”, and so on. But we do not believe that this suggestion can account for our data. Because we only considered the 100 most frequent words, few were highly domain specific: in fact, only “theorem” and “proof” were words in the overall top 100 which had fewer than 100 occurrences in the Brown/LOB corpus.

Instead our conclusion is that the undergraduate students who provided the texts for our learner corpus did successfully produce written mathematics that was consistent with that found in undergraduate textbooks written by expert mathematicians. At least in the sense that it shared a similar distribution of word frequencies.

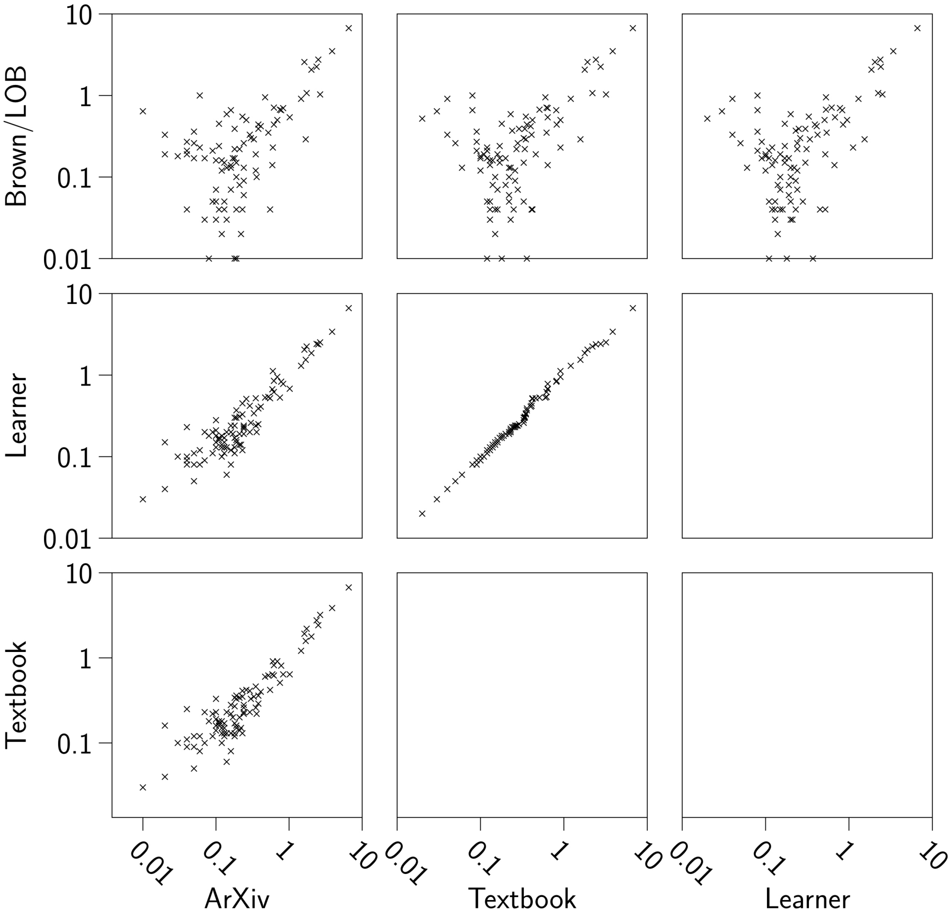

Figure 1. Scatterplots showing the frequencies of the top 100 words (as percentages) for each pairwise combination of our four corpora. Axes have logarithmic scales (therefore words with zero frequency in one corpus are not shown).

The main difference we found between the mathematical corpora concerned the word frequency distributions of the textbook and arXiv corpora. We can explore this difference in more detail by considering the keywords for each corpus – those words which occur disproportionately in one corpus compared to the other. These are shown in Table 2, which is ordered by chi-squared value (i.e. the contribution of the word to S_{AB} defined earlier).

Some of these key words are unsurprising: for instance, 'example', and 'solution' occur proportionately more often in the textbook corpus than the arXiv corpus. The textbook corpus also contains proportionately more instances of verbs such as 'find', 'show', 'do', and 'prove' – than the arXiv corpus. Indeed, the only verbs appearing in the right-hand side of Table 2 are 'let' and 'see'. Although one might attribute this to the inclusion of exercises in textbooks, this explanation would not account for the extremely similar frequencies for these words found in the textbooks and the undergraduates' final-year projects: although clearly textbooks normally contain exercises, student projects do not.

One further difference between the mathematical corpora concerned the frequency of mathematical notation. The arXiv corpus had considerably more instances of "inline_math" per 100 words (11.1%) compared to the textbook (8.8%) or learner (7.6%) corpora.

Further analyses are required to understand the significance of some of the other differences between the corpora. For instance, 'by' occurs disproportionately often in the arXiv corpus (1.02% of words) compared to the textbook corpus (0.64%) , but why? Investigating the most common clusters of words that include 'by' in the arXiv corpus indicates that the word is used to both name ("defined by", "denote by") and assert ("given by", "obtained by", "generated by"). By systematically studying such cases we can begin to understand the differences between research-level and undergraduate-level mathematical language.

Conclusion

Our main goal in this paper has been to describe the creation of three mathematical corpora designed to aid researchers understand mathematical language. The tools we used to construct these corpora are freely available for the research community to use. Having constructed the corpora we presented an analysis of word frequency distributions which suggested that undergraduate students are, by the end of their courses, surprisingly successful at writing in a manner consistent with the language used in undergraduate textbooks. The developmental trajectory by which students develop mathematical language skills would be a worthy topic of future study. In contrast to the similarity observed between textbooks and final year dissertations however, the language mathematicians use in research papers is different to both.

In this paper we have focused on comparing the word frequency distributions of four different corpora, but there are a great many other techniques that can be used to analyze corpora which go well beyond this approach (e.g., McEnery & Hardie, 2011). Given the interest shown by mathematics educators and other researchers in mathematical language, we believe that corpus linguistics is a potentially useful, but currently under used, research technique.

More frequent in the textbook corpus	
Word	χ^2
find	0.01279
what	0.01103
number	0.00689
example	0.00477
if	0.00266
must	0.00265
use	0.00265
show	0.00242
function	0.00232
set	0.00220
that	0.00218
or	0.00218
about	0.00212
is	0.00201
solution	0.00181
do	0.00167
two	0.00163
not	0.00159
each	0.00141
than	0.00139
so	0.00130
prove	0.00124
this	0.00123
are	0.00113
a	0.00113

More frequent in the arXiv corpus	
Word	χ^2
by	0.00367
on	0.00190
where	0.00144
case	0.00120
i	0.00120
with	0.00100
for	0.00094
let	0.00071
in	0.00061
see	0.00058
following	0.00055
such	0.00054
space	0.00050
which	0.00049
proof	0.00033
theorem	0.00028
also	0.00019
we	0.00015
now	0.00013
given	0.00010
our	0.00007
and	0.00007
since	0.00003
from	0.00002
every	0.00002

Table 2: The left-hand table shows the top 25 words that occur in the textbook corpus that differentiate it from the arXiv corpus. The right-hand table shows the equivalent words for the arXiv corpus.

References

Francis, W. N. & Kucera, H. (1961). *Brown Corpus Manual.* Providence, RI: Brown University.

Inglis, M., Mejía-Ramos, J. P., Alcock, L. & Weber, K. (2013). On mathematicians' different standards when evaluating elementary proofs. *Topics in Cognitive Science, 5,* 270-282.

Johansson, S. (1986). *The tagged LOB Corpus: User's Manual*. Bergen, Norway: Norwegian Computing Centre for the Humanities.

Larvor, B. (2016). *Mathematical Cultures: The London Meetings 2012-2014.* Basel, Switzerland: Birkhäuser.

McEnery, T. & Hardie, A. (2011). *Corpus Linguistics: Method, theory and practice.* Cambridge: CUP.

Stylianides, A. J. (2007). Proof and proving in school mathematics. *Journal for Research in Mathematics Education, 38,* 289-321.

Vivaldi, F. (2014). *Mathematical Writing*. London, UK: Springer.