Corpus-based teaching materials for specialised courses: the case of English for mathematicians

Lucie Malá

Abstract: This paper demonstrates a usage-based approach, drawing on language corpora, to creating teaching materials. The need for such materials stems from the growing demand for courses of English for specific purposes on one hand, and the surprising lack or inadequacy of available materials for these courses on the other hand. On the example of English for mathematicians, we present three different types of exercises. First, we show deductive exercises for practising an already familiar rule. Second, we focus on the creation of an exercise where students start with a general rule but derive more specific information from specifically chosen examples. Finally, we present a fully inductive exercise, aimed at familiarising the students with the structures specific of the texts from their discipline. We believe that these exercises may serve as a source of inspiration for other teachers of English for specific purposes who find themselves in a similar situation.

Key words: ESP, teaching materials development,corpora, mathematics

1 Introduction

In this paper, we demonstrate the use of corpora in developing teaching materials for a specialised course of English for mathematicians. The use of corpora, i.e. collections of natural occurring texts (Hunston, 2002), in teaching and learning English for Specific Purposes (ESP) is becoming more and more widespread. Typically, three broad roles of corpora are mentioned: corpora as basis of research underpinning the contents of ESP courses, corpora as tools that the learners themselves exploit in acquiring the target language, and finally corpora as resources for teachers' classroom activities and materials (Hewings, 2012). These areas have been explored in a growing number of studies, summarised for example in Boulton et al. (2012), Gavioli (2005), and Timmis (2015). While many of these and other studies provide ideas of creating corpus-based materials, we hope to bring a discipline specific point of view which might be inspirational for our readers.

2 The need for specialised materials

Before introducing the corpus-based ESP materials, we justify the need for such materials, arguing that their creation is not only desirable, but also necessary.

A vast body of research into disciplinary discourses (see e.g. Gray, 2015; Hyland, 2004, 2011), i.e. the ways language is used by academics of a particular discipline, has pointed out differences between individual disciplines. In recent years, this research activity has resulted in an increase in demand for specialised courses

of English, which no longer focus on academic English in general, but on English as it is used by practitioners of a particular discipline. Accordingly, these courses should be based on specialised teaching materials.

English for mathematicians suffers from two main problems regarding course materials. First, there are virtually no teaching materials concerned with mathematical English. The existing textbooks are limited to writing style manuals (e.g. Higham, 1998), often focused on typographical conventions more than on specific textual features. None of the materials is a textbook as such, suitable for direct use in the classroom[1]. Moreover, all of the available materials are fairly outdated, being published around the year 2000 at best.

Second, mathematical English is to a large degree distinct from general academic English (EAP). In the absence of a specialised textbook, resorting to general EAP materials, which are abundant, would seem an obvious solution. However, looking at the topics traditionally covered in the EAP courses, we see that the language of mathematics differs quite radically from what is understood to be academic English. As an example, we mention three characteristic features of academic writing usually discussed in EAP materials (e.g. J. M. Swales & Feak, 2012). First, students are encouraged to use ways of moderating and qualifying a statement, generally referred to as hedges. However, it has been shown by McGrath and Kuteeva (2012) that the number of hedges in mathematical writing is extremely low compared to other disciplines, and some forms of hedging have not been attested at all. Second, it is a common practice to teach students to use passive voice in their writing, especially in describing processes. In contract to this, mathematical writing abounds in *we* and consequently active voice (ibid.), especially in describing procedures. Finally, the macro-organisational structure of mathematical research papers does not correspond to well-established models such as Introduction-Method-Results-Discussion (J. Swales, 1990). Some sections are completely absent, while other attested sections seem to be specific of mathematics (Graves et al., 2013, 2014; Moghaddasi & Graves, 2017). In conclusion, general academic materials are unsuitable for a course of English for mathematicians.

3 Material

To create quality corpus-based materials, it is necessary to have access to a corpus representative of the variety of language that is taught in the ESP. Nowadays, a number of corpora are freely available online. Unfortunately, there is only one corpus of mathematical writing that we know of, and it was not constructed to be representative of this discipline, as it is a part of a more general corpus

[1] This is with the one exception of *Angličtina (nejen) pro studenty MFF UK* (Křepinská et al., 2019), which is however aimed at general preparatory classes at the bachelor level.

of published journal papers across disciplines (Kosem, 2010). The exercises are therefore based on our own corpus of mathematical research papers[2] (CoMaR) designed to be representative of this variety and reflect the specialisations of the target students. The corpus consists of 108 research papers, amounting to 870,885 words.

In this paper a licensed online corpus analysis tool SketchEngine (Kilgarriff et al.) is used to obtain data from the corpus, but the exercises are not dependent on any specific functionality offered by this tool and could have been created using freely accessible software, e.g. AntConc (Antony, 2019), or LancsBox (Brezina et al., 2020).

4 Corpus-based exercises

In this section we present three types of exercises based on our corpus. We start with an exercise which fits the deductive approach to teaching, and move along the deductive-inductive scale towards a strongly inductive approach (as defined in e.g. DeKeyser, 1995, p. 380).

4.1 Example 1: Articles

We start with an exercise which raises awareness of certain phenomena the students are familiar with from general English but whose usage differs slightly in specialised English for mathematicians. In this particular case it is the use of articles. Namely, students practise rules for the use of zero article with countable nouns, which is in mathematics common in the context of numbered items, e.g. *Theorem 1*, names of mathematical disciplines, e.g. *number theory*, and characteristics of objects after *with* or *have*, e.g. *a circle with centre O*, *the matrix has rank 2*.

As can be seen in Figure 1, students are asked to cross out all unnecessary articles based on the set of rules they were given. In this case, the corpus serves as a source of representative example sentences. To find these, we use a simple corpus search and extract all sentences that include a selected target word or expression, e.g. *number theory*, or *centre*. The teacher's task is then to select those sentences which are most representative and easiest to understand for the group of students. Choosing the search words carefully, it is also possible to create an exercise that fits the current thematic focus of the class, e.g. geometry, or linear algebra.

[2] This corpus has been designed for the purposes of the author's PhD thesis. Details of its design and construction will be published in the thesis *Mathematical texts from the perspective of distributional phraseology* (forthcoming).

Study the rules for zero articles with countable nouns in mathematics. Then cross out all extra articles.

a) *Finally, simple generalizations of some of the concepts in the number theory to integer square matrices are presented.*

b) *Let a graph* $G = (V, E)$ *be defined by a set* $V = \{0, \ldots, N-1\}$ *of vertices, with the cardinality* $|V| = N$*, and a set* E *of edges.*

c) *The Figure 2 shows some spectra of the prime sets from the Example 6.4 with the lengths 3, 4, and 5, respectively; and with the widths 2, 2 and 1, respectively.*

d) *In Rn, we denote by* $B = B(x, r)$ *an open ball with the center* x *and the radius* $r > 0$*.*

e) *Suppose the conditions in the Propositions 4.1 and 4.2 are satisfied, and* V *has the rank* d*.*

Fig. 1: *An example of a deductive corpus-based exercise*

4.2 Example 2: Universal quantifier

This exercise answers the question how the universal quantifier is expressed in words. In contrast to the first example, the corpus serves not only as a source of examples but also as a source of information for the teacher. Our initial intuition would be that the universal quantifier is expressed by *for all* and *for every*. We will first try to confirm this conjecture. To do so, we use a corpus search with a wildcard, i.e. searching for "for *", where the asterisk can stand for any word. This form of search stems from the assumption that while the realisation of the grammatical quantifier can vary, the preposition remains the same. The concordance lines obtained through this search will not be of much use, as phrases such as *for computer*, *for example*, and *for convenience*, will be included. Fortunately, the use of the universal quantifier can be safely assumed to be quite frequent in mathematical texts. Therefore, we can look at the frequencies of the distinct phrases found by our search. An excerpt of this table based on a search in our corpus can be seen in Figure 2.

From expressions found on top of the list it seems that *for all*, *for any*, *for each*, *for every* could correspond to the meaning of the universal quantifier. It is of course necessary to go back to the concordance lines for each of these expressions to verify that it really functions as a universal quantifier.

Having done this, we can prepare an exercise which will familiarise the students with these various expressions and the contexts in which they are used. Once again, the teacher needs to select the most suitable examples from the corpus as a basis of the exercise. It is the advantage of using a whole corpus for drawing

	Lemma	↓ Frequency	Per million tokens
1	for the	1,654	1,011.84
2	for all	1,017	622.15
3	for any	755	461.87
4	for some	561	343.19
5	for a	517	316.28
6	for each	459	280.79
7	for every	397	242.87
8	for example	231	141.31
9	for k	179	109.50
10	for i	150	91.76
11	for n	127	77.69

Fig. 2: *The first eleven most common hits for the search "for *" in CoMaR.*

these examples that the teacher can get a more precise idea of what a prototypical use of the phrase is. They are then in a better position to select examples highlighting the phrases' typical characteristics.

Figure 3 shows an exercise created in the above described way. Students are asked to complete the statements with the target expressions. The teacher asks the students how they would express the phrases in symbols, to make them realise that they all correspond to the universal quantifier. The gap-filling is in this case a starting point for a deeper analysis of the sentences and discussion. The discussed questions include what would change if another of the expressions was used, if some of them are interchangeable, what the preferred position and context for each of the phases is, and how to translate them.

While students will need the teacher's help in answering some of the questions, they are able to provide answers to others independently, based on the sentences. We can therefore say that this exercise is somewhere in between the deductive and inductive approaches. The students start with a rule of how the universal quantifier is translated, but derive more details about its use through examples.

4.3 Example 3: Presentative *let*

The last example is an inductive exercise where students uncover a particular construction typical of mathematical texts and describe it. The construction in question can be schematically written as *let* [symbol]*be* [noun phrase]. A simple example is *let G be a group*. This is, in fact, an instantiation of a more general construction, where other verbs might be used, and the noun phrase can be replaced by an adjective or adjectival phrase, e.g. *let x denote the axis of symmetry*, *let p be positive*. However, this construction is difficult to use for our students and

Complete the following statements with *for all* / *for each* / *for every* / *for any*.

a) Lemma 2.8. (Bezout's identity). _____ integers $a, b \geq 1$, there exist integers m, n such that $ma + nb = \gcd(a, b)$, where $\gcd(a, b)$ denotes the greatest common divisor of a and b.

b) Definition 1. A function: $\phi : KmR \rightarrow R$ is said to be superquadratic if _____ $x \in Km$ there exists a vector $c(x) \in Rm$ such that (1.1) $\phi(y) \geq \phi(x) + c(x)$, $y - x + \phi(|y - x|)$ holds _____ $y \in Km$.

c) Lemma 6.4. (Stable Decomposition). _____ $v \in V = V^h(\Omega)$, there exists a decomposition $u = \sum_{i=0}^{N} R_i^T u_i$ with $u_i \in V_i = V^h(\Omega_i')$ such that [formula].

d) Definition 4.1. We say that a set B additively generates another set A if _____ $a \in A$ there exists some subset $\{b_1, b_2, \ldots, b_n\}$ such that $a = \sum_{i=1}^{n} b_i$.

e) Lemma 4.2. There exists $p < 0.5$ such that a projection P_j satisfies $||P_m P_{m+1} \ldots P_{n-1}|| \leq C 2^{p(n-m)}$ _____ $2 \leq n < n$ and a constant C independent of m and n.

Fig. 3: *An example of a deductive-inductive corpus-based exercise*

experience shows that it is better to first focus on the narrower structure and then to generalise.

As with the other examples, we first need to obtain sentences containing this construction. There are several ways of approaching this. It is possible to use a simple search for "let", and then filter the lines that contain *be* in the right context of the search word. Alternatively, if the chosen corpus analysis tool allows this, the query can be formulated in the Corpus Query Language (CQL). In this case we can search for a sequence of the lemma *let*, i.e. *let* in any form including *letting*, or *Let*, followed by one to five further unspecified tokens, and the word *be*. The CQL is then:

[lemma="let"] []{1,5} [word="be"]

If the corpus is tagged for word classes, it is possible to further specify that we are looking only for cases where the described sequence is followed by a noun at some distance:

[lemma="let"] []{1,5} [word="be"] []{0,4} [tag="N.*"]

For our purposes of obtaining examples of use, being overly specific is not only unnecessary, but also often undesirable as it might accidently eliminate relevant examples due to inaccurate formulation of the query, or mistakes in tagging. With

a slightly wider search the teacher will only need to ignore irrelevant sentences. We have used the first query to obtain material for this example exercise.

A single word is missing from the following sentences. What is it?

1			x and y be nodes of a binary tree.
2	For $1 \leq a < b < c \leq t$,		l_{abc} be the line containing x_{ab}, x_{ac}, x_{bc}.
3	... and		$\sim$ be the least multiplicative equivalence upon a free loop F (X) for which $x \sim y$.
4			k be an arbitrary field.
5			u and v be distinct variables not occurring in ϕ.
6	For every $h \in V$ with $h \equiv_\Gamma f$,		h^* be the sequence given as follows: $h^*(i) = a, h^*(j) = b$ and $h^*(k) = h(k)$, for every $k \in \alpha \sim \{i, j\}$.
7			Ω be an open polygon in R^2 or an open Lipschitz polyhedron in R^3, with Lipschitz boundary $\Gamma := \partial\Omega$.
8			Ψ and Σ be bounded sets of positive kernel operators on a Banach function space L.
9			$p \subset M$ be a finite prime set.
10			$\hat{v}_1, \ldots, \hat{v}_d$ be the eigenvectors of the matrix $\widehat{V}_n$ corresponding to its d largest eigenvalues.

Fig. 4: *An example of an inductive corpus-based exercise*

Students are presented with a set of sentences representative of the target construction. Notice that in this case the sentences are arranged in the same way as in the corpus KWIC (=key word in context) view, with the key item in the central column. The place where this word should be is highlighted in yellow in Figure 4. This arrangement is especially useful for noticing similarities in the structure of the sentences. It is the first task of the students to guess this one missing word, i.e. *let*. This is merely an introductory task which makes the students read the sentences, and reminds them of the structure, which they have seen before.

The main part of the exercise comes after the word *let* has been filled in. Students focus on the form of the sentences, try to describe similarities and generalise the formal structure of the underlying construction. They are then asked to think about the influence of the concrete realisation of the noun phrase. These are possible questions they might be asked:

- Is the noun part in singular or in plural? Does this change anything?
- Which articles can be used in the noun phrase part of the construction? What does the choice depend on? Which article would you predict to be most commonly found in this construction?

The understanding of the construction is based on the sentences provided, but also on students' experience and intuitions about mathematical meanings, as is apparent in the last question. This is even more pronounced in questions about the function and position of the construction within mathematical texts, such as:

- What is the function of the structure? Why is it used in mathematical texts?

- How would you translate it into your native language?
- Where can it be found with respect to sentences? ...paragraphs? ...the whole mathematical texts?

The ability to answer such questions depends on the amount of experience of the students. The teacher might want to make these easier by including wider context of the construction, or a greater number of examples. In the case of 'presentative let' adding more left context would reveal that it usually stands at the beginning of paragraphs, and typically at the beginnings of specific mathematical sections, e.g. a proof, definition, or a theorem.

With students with a wider experience with mathematical writing this exercise can be expanded by asking about what other elements might fill the place of the noun phrase, or what other verbs they have seen in the construction.

The role of the corpus in creating this exercise is double fold. First, it serves as a source of authentic examples. Second, it can enhance the teacher's understanding of the use of the construction and help provide answers to the above suggested questions. In this type of exercises, it sometimes happens that students make an observation the teacher has not expected. The corpus then provides a means of verifying the students' hypothesis.

5 Conclusions

Teaching English for specific purposes of a particular discipline entails the challenge of finding suitable classroom materials. We have explained that for mathematics in particular this task is made more difficult still by the surprising lack or inadequacy of the existing textbooks and the impossibility of using EAP materials. The use of corpora in developing specialised materials is then seen as a convenient solution. It provides a usage-based approach, drawing on naturally occurring language in context. Moreover, the teacher can create exercises corresponding to students' needs, as well as the course syllabus. Finally, this paper demonstrated that it is possible to accommodate corpus-based exercises to both deductive and inductive approaches to teaching.

References

Anthony, L. (2019). AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: WasedaUniversity. Available from http://www.antlab.sci.waseda.ac.jp/

Boulton, A., Carter-Thomas, S., & Rowley-Jolivet, E. (Eds.). (2012). *Corpus-informed research and learning in ESP: Issues and applications*. John Benjamins Pub. Co.

Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBoxv. 5.x. [software]. Available at: http://corpora.lancs.ac.uk/lancsbox

DeKeyser, R. M. (1995). Learning second language grammar rules: An experiment with a miniature linguistic system. *Studies in Second Language Acquisition*, *17*(3), 379–410.

Gavioli, L. (2005). *Exploring corpora for ESP learning*. John Benjamins.

Graves, H., Moghaddasi, S., & Hashim, A. (2013). Mathematics is the method: Exploring the macro-organizational structure of research articles in mathematics. *Discourse Studies*, *15*(4), 421–438. https://doi.org/10.1177/1461445613482430

Graves, H., Moghaddasi, S., & Hashim, A. (2014). "Let G = (V, E) be a graph": Turning the abstract into the tangible in introductions in mathematics research articles. *English for Specific Purposes*, *36*, 1–11. https://doi.org/10.1016/j.esp.2014.03.004

Gray, B. (2015). *Linguistic Variation in Research Articles*. John Benjamins Publishing Company.

Hewings, M. (2012). Using Corpora in Research, Teaching, and Materials Design for ESP: An Evaluation. *Taiwan International ESP Journal*, *4*(1). https://doi.org/10.6706/TIESPJ.2012.4.1.1

Higham, N. J. (1998). *Handbook of writing for the mathematical sciences* (2nd ed). Society for Industrial and Applied Mathematics.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.

Hyland, K. (2004). *Disciplinary Discourses: Social Interactions in Academic Writing*. The University of Michigan Press.

Hyland, K. (2011). Disciplinary specificity: Discourse, context and ESP. In D. Belcher, A. M. Johns, & B. Paltridge (Eds.), *New Directions in English for Specific Purposes Research* (pp. 6–24). University of Michigan Press.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2021). The Sketch Engine. http://www.sketchengine.eu

Kosem, I. (2010). *CAJA: Corpus of Academic Journal Articles|Sketch Engine*. https://www.sketchengine.eu/corpus-of-academic-journal-articles-caja/

Křepinská, A., Bubeníková, M., & Mikuláš, M. (2019). *Angličtina (nejen) pro studenty MFF UK* (2.). MatfyzPress.

McGrath, L., & Kuteeva, M. (2012). Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. *English for Specific Purposes*, *31*, 161–173.

Moghaddasi, S., & Graves, H. A. B. (2017). "Since Hadwiger's conjection… Is still open": Establishing a niche for research in discrete mathematics research article introductions. *English for Specific Purposes*, *45*, 69–85.

Swales, J. (1990). *Genre analysis: English in academic and research settings* (13. printing). Cambridge Univ. Press.

Swales, J. M., & Feak, C. B. (2012). *Academic Writing for Graduate Students: Essential Tasks and Skills* (3rd edition). The University of Michigan Press.

Timmis, I. (2015). *Corpus linguistics for ELT: Research and practice*. Routledge, Taylor & Francis Group.

Author

Mgr. Lucie Malá, e-mail: malaluci@mbox.troja.mff.cuni.cz is an English teacher at the department of language education at the faculty of mathematics and physics of the Charles University. Having studied both English and mathematics, she specialises in teaching English for mathematicians. She is responsible for the organisation of UNIcert® III, English for Mathematicians teaching programme and examination administered at the faculty. Lucie is also a fourth year PhD student of English at the Faculty of Arts, where her topic of research is phraseology of mathematical research articles.