

Citation for published version:

Ratkaine Jablonkai, R & Cebon, N 2021, Undergraduate students' responses to a corpus-based ESP course with DIY corpora. in M Charles & A Frankenberg-Garcia (eds), *Corpora in ESP/EAP Writing Instruction: Preparation, Exploitation, Analysis*. Routledge, London, U. K., pp. 100-120.

Publication date:
2021

Document Version
Peer reviewed version

[Link to publication](#)

University of Bath

Alternative formats

If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Please cite as:

Jablonkai, R. R., & Čebroň, N. (2021). Undergraduate students' response to a corpus-based ESP course with DIY corpora. In M. Charles & A. Frankenberg-Garcia (Eds.). *Corpora in ESP/EAP writing instruction: Preparation, Exploitation, Analysis*. Routledge.

Undergraduate students' response to a corpus-based ESP course with DIY corpora

Reka R. Jablonkai and Neva Čebroň

Abstract

To date there has been little research into how do-it-yourself (DIY) corpora can effectively be used in DDL courses in professional ESP contexts with undergraduate students. This chapter reports on the design, outcomes and value of a course integrating corpus compilation and consultation for undergraduate students of ESP. The chapter introduces some core principles of course design for corpus-based ESP courses, arguing that approaches proposed by CALL research should be incorporated in corpus-based courses. Data was collected by means of analysis of written student reports and questionnaires administered immediately after the course and five months after the end of the course. Findings reveal that all students found corpus use beneficial for ESP learning. They especially valued corpora for writing, expanding their vocabulary in general and learning subject-specific terminology. A third of the students mentioned specific plans to use corpora in the future in their reports and a great majority of students reported having used online available corpora five months after the course. It is suggested that the process of compiling their DIY corpora may have enhanced their understanding of the nature and limitations of corpus data and thus helped them interpret corpus output more accurately.

Introduction

The last couple of decades have seen a surge in research into corpora in language teaching and in particular in teaching writing (e.g. Boulton & Cobb, 2017; Chen & Flowerdew, 2018a). To date, however, little research has been conducted on integrating corpus consultation into undergraduate ESP courses, with the majority of studies focusing on EAP courses for post-graduate and doctoral students (Chen & Flowerdew, 2018a). The identified benefits of corpus use include an increase in learners' language awareness, confidence and autonomy (Chen & Flowerdew, 2018a; Yoon, 2008). Corpora have been used especially successfully in EAP writing courses where the aim was to demonstrate discipline-specific discourse and lexicogrammatical features of academic genres (e.g. Charles, 2012; Lee & Swales, 2006). As Yoon (2008) suggests "if well planned, learning with corpora can enhance the learners' writing experiences and their confidence" (p. 46). While studies have primarily investigated the effects of corpus-based and data-driven learning (DDL) courses, little attention has been paid to what makes a "well planned" corpus-based course. At the same time, there are few reports on the application of DDL in ESP contexts where the aim is to sensitize students to subject-specific variation in linguistic features for their future profession. Therefore, the present study focuses on the design, outcomes and value of a corpus-based ESP course for undergraduate students. It draws on two research strands: firstly, on studies on corpora in language teaching,

and secondly, on research in the field of computer-assisted language learning (CALL). Although the results are modest due to the small number of participants, they point to the essential elements and approaches relevant to corpus-based ESP courses for undergraduates.

Course design for corpus-based ESP courses

The central idea of DDL and corpus-based courses is to provide learners with resources and skills in the use of corpora so that they become autonomous (Johns, 2002), that is, they take control over their own language learning (Holec, 1981). While there are numerous studies on the effects of DDL courses (Boulton & Cobb, 2017; Chen & Flowerdew, 2018a), only a handful of studies have concentrated on the design and content of these courses (Leńko-Szymańska, 2014; 2017; Yoon & Jo, 2014). These studies primarily investigated courses for future language teachers and aimed to evaluate the effectiveness of teacher training courses in terms of students' ability to use corpora autonomously as language learners and for pedagogic purposes (Callies, 2016; Frankenberg-Garcia, 2012; Leńko-Szymańska, 2014; 2017; Hüttner, Smit, & Mehlmauer-Larcher, 2009; O'Keeffe & Farr 2003). The three skills and competencies that were identified as relevant to be developed in DDL courses are: 1) technical skills, that is, students need to be computer literate and competent in using corpus resources and tools; 2) an understanding of corpus linguistics, that is, students should be able to carry out corpus searches and to interpret results; and 3) pedagogic skills, that is, students should be able to implement corpus data in their own teaching practice (Callies, 2016; Leńko-Szymańska, 2017).

In order to develop technical skills, Leńko-Szymańska (2014) recommended introducing students to a range of corpora and corpus tools so that they do not become dependent on the single tool they were trained to use. Corpus-based courses often use large, general corpora, for example, the BNC and COCA (e.g. Bridle, 2019; Yoon, 2008; Yoon & Hirvela, 2004; Yoon & Jo, 2014) and less frequently parallel corpora (Frankenberg-Garcia, 2005). In EAP and ESP courses, small discipline-specific corpora are often used for comparison purposes or to demonstrate textual and lexico-grammatical features of specific subject fields, registers or genres (Hüttner et al., 2009; Ackerley, 2017; Chen & Flowerdew, 2018a). For example, Ackerley (2017) prepared a small corpus of public opinion survey reports to be used in her DDL classes as an expert corpus that served as a model for her students. In a few studies students were asked to compile their own do-it-yourself (DIY) corpora to be used as discipline-specific models for their writing or their future professions (Charles, 2014; 2017; Godwin-Jones, 2017; Jablonkai & Čebon, 2017; Lee & Swales, 2006; Leńko-Szymańska, 2017). There is, however, very little research on how self-compiled corpora can successfully be used with undergraduate ESP students.

To give students practice in interpreting corpus search results, DDL courses apply two main approaches: direct and indirect corpus use. In the case of direct corpus consultation, learners use corpus software to access corpora and explore lexico-grammatical patterns themselves (Chambers, Farr, & O'Riordan, 2011; Pinshuan & Lin, 2019). In contrast, indirect corpus use refers to paper-based, carefully edited concordance lists presented to students in the form of handouts (Yoon & Jo, 2014). As regards the benefits of direct and indirect corpus use, Yoon and Jo (2014) found that students were more successful in error correction with the indirect

use of corpora. However, when they looked at students' preferences, there was a clear preference for direct corpus use. Other studies reported using a mixture of indirect and direct corpus use in their courses (Chen & Flowerdew, 2018b; Ackerley, 2017). In these studies, printouts of corpus searches were used in the initial stages of the course and direct corpus use was introduced later. The specific DDL tasks primarily centred around interpreting frequency lists, analysis of lexico-grammatical features of specific vocabulary items and observation of language use and rhetorical functions in specific text types (Ackerley, 2017; Charles, 2018; Chen & Flowerdew, 2018a; Frankenberg-Garcia, 2018).

A useful technique applied in a handful of studies to enhance students' engagement in direct corpus use is to contextualise corpus consultation activities so that in-class corpus use demonstrates how corpora can be used for students' future purposes of corpus consultation. Charles (2018), for example, showed how she introduced specific tools to students by not only demonstrating how they are operated, but also how they can be specifically used for editing purposes which was the most likely purpose for which her doctoral students would use corpora in the future.

Learner training in CALL

In his systematic review of DDL literature, Pérez-Paredes (2019) points out that DDL is often not perceived as CALL-related research, probably because its origins lie in language rather than technology-oriented research. It should also be added, that there seems to be little uptake of CALL research in DDL literature. The few exceptions include, for example, Friginal (2018), who devotes a chapter to discussing how corpus-based approaches to language teaching can be situated within CALL. There are, however, important lessons DDL courses can learn from CALL research, especially when it comes to course design and learners' preparation for corpus consultation and compilation.

As regards preparing learners to using CALL technology, Hubbard (2013) argues for learner training as an important element. He proposes a framework of practice-based learner training principles to enhance autonomous use of CALL technology. He suggests that teachers take a learner perspective and at the same time provide learners with pedagogical knowledge so that they are well-informed to plan their own CALL activities. When learners acquire new skills and knowledge, Hubbard recommends a cyclical approach with repetitions so that learners have time for hands-on practise and to internalise the necessary skills. He emphasises the importance of collaborative debriefings to monitor learners' progress and as a social learning opportunity for learners to reflect on their use of CALL technology.

Most DDL courses include some of these elements, for example, time allocated to hands-on practice (e.g. Ackerley, 2017; Charles, 2012), however, little attention is paid to other principles, for example, reflection and collaborative debriefing. The present study aims to integrate these learner training principles into the course design of a corpus-based ESP course.

Tentative guiding principles for corpus-based ESP course design

The review of the DDL and CALL literature points toward some tentative guidelines for DDL courses. Effective corpus-based ESP courses:

1. Introduce students to basic concepts of corpus linguistics.
2. Include components that focus on technical skills and demonstrate a selection of corpus resources and tools.
3. Introduce students step by step to corpus consultation by, for example, including both direct and indirect corpus use or guided and scaffolded corpus consultation.
4. Allow sufficient time and repetition for hands-on practice corpus querying in class.
5. Include compiling and analysing a subject-specific DIY corpus.
6. Promote autonomous corpus use outside of the classroom.
7. Contextualise the demonstration of corpus resources and tools.
8. Include components of reflection and debriefing.

These guidelines were followed in the course design in the present study to develop and deliver a corpus-based ESP course for undergraduate students. The study examined the course outcomes in terms of the results of students' corpus consultation, their reflection on and perception of corpus use. The following research questions were formulated to guide the study:

- 1) In what ways do undergraduate students think corpus use is beneficial for their ESP learning?
- 2) What are the perceived challenges of corpus use among undergraduate students of ESP?
- 3) What are undergraduate ESP students' perceptions of the compilation and consultation of DIY corpora as reflected in their written reports after a corpus-based ESP course?

Institutional context and participants

The Corpus-based ESP Course reported here is part of the undergraduate study programme on Intercultural Linguistic Mediation at the Faculty of Humanities of the University of Primorska, Slovenia. It is run as a mandatory course for undergraduate students in their second year of studies, while also offered as an optional course to students from other study programmes. Students in this course study English as a major and choose between Italian and French as a minor.

The course is delivered throughout a whole semester of 15 weeks. It is held as weekly 5-hour sessions in a computer laboratory with groups of 10 to 15 undergraduate students. The participating students are required to have a good level of English (upper-intermediate to advanced) and basic computer literacy (e.g. Internet searching, word processing skills). The aim of the course is to prepare students to work as language professionals in various multilingual settings as language editors, language service providers or translators.

Rationale for a DDL approach

It is envisaged that students' future working environment will demand from them the ability to deal with a number of situations where various types of specialist language knowledge could be required. Therefore, it was considered essential that during their undergraduate studies students not only acquire a high level of proficiency in general English, but also develop skills and competencies enabling them to research language use in various genres and disciplines so as to be able to identify the salient features of individual language varieties.

As previous studies demonstrated that DDL courses result in enhanced language awareness (Chen & Flowerdew, 2018a; Yoon, 2008) and provide students with tools to identify discipline-specific linguistic features (Gavioli, 2005; Crosthwaite & Cheung, 2019), a DDL approach was considered appropriate for this purpose.

The core aims of the Corpus-based ESP course can be summarized as follows:

- develop students' language skills for their future professions
- sensitize students to variation in lexico-grammatical features of subject-specific language use
- raise awareness of the formulaic quality of naturally occurring language
- enhance autonomous language learning
- provide a tool for students' long-term professional use

Participants

This study investigated a group of 9 students who attended the Corpus-based ESP Course in the 2018/19 academic year. Participants ranged in age from 20 to 31 and were all female. By nationality, the group comprised 8 Slovenians and one Turkish student. Their level of English was assessed with a placement test prior to enrolling on the course and ranged from intermediate (3 students) to advanced level (6 students). Answers to a pre-course questionnaire revealed that the majority of participants had heard about corpora (7 students), but only a few (3 students) had used available corpora, for example, English language corpora: COCA or the BNC; Slovene language corpus: GigaFida, before taking the course.

Course design and procedure

The main tenet followed in designing the course was that at the early stage of undergraduate studies students could profitably engage in corpus-based activities in order to expand and refine their language skills through a DDL approach (Crosthwaite & Cheung, 2019; Johns, 1991). The course design followed the guiding principles established on the basis of previous studies. The corpus-based approach allowed the course to utilise the benefits of discovery learning (Chambers, Farr, & O'Riordan, 2011) instead of explicitly teaching subject-specific vocabulary and lexico-grammatical features. An important guiding principle for the design and delivery of the course was, therefore, to allocate sufficient time for students to gradually discover corpus consultation approaches through tasks and exercises, which allowed them to focus on form, i.e., observation of words, collocations, grammatical structures, pragmatic patterns in context (Long, 1991). Furthermore, building on Hubbard's (2013) principles, we embedded classroom training in a progressive, cyclical manner with small segments of training content in each session and concluding sessions with collaborative debriefings.

In order to gradually develop students' necessary technical skills and to introduce them to basic concepts of corpus linguistics, the course was divided into the following four modules (see also Appendix 1): 1) the introductory module dealt with core concepts of corpus-based analysis and resources; 2) the second module focused on principles of corpus compilation; 3) the third module introduced the analysis of DIY corpora (frequency lists, keywords, n-grams and ESP terminology); 4) the final module provided practice in corpus consultation for written production (essays or translations). All the sessions consisted of a short presentation of the topic with theoretical and practical insights delivered by the teacher, while most of the

time was dedicated to hands-on practical work structured around the tasks given in worksheets.

The first module introduced students to basic concepts of corpus linguistics and corpus-based research. Consultation of corpus data was demonstrated by exploiting on-line corpora in both English and Slovene, namely COCA and GigaFida. Contrasting statistical patterns of contextualised language in use, including data on word frequency, collocation and colligation in both languages helped scaffold students' understanding of variation in lexical patterns within a language and across languages. The formulaic quality of naturally occurring language (Altenberg, B., 1998; Wray, A., 1998) was observed and analysed.

In the second module, students were guided to building their own subject-specific corpus, opting for any topic of their interest. To introduce the main principles of corpus compilation, two pre-compiled corpora were analysed in terms of structure and balance of documents. These were a corpus of Annual Reports of various multinational companies and a corpus of texts on Cultural Heritage, both around 1,000,000 running words. These were used because they display radically different varieties of language use and served as models to introduce core corpus compilation principles (e.g. domain, genre, register, balancing text types), which need to be considered in the selection and gathering of documents for small self-compiled specialized corpora. The guiding principle in text selection was to focus narrowly on a specific domain so that the lexico-grammatical patterns characteristic of their specific disciplinary corpus would be more prominent. From this stage on students started working exclusively with their DIY corpora.

The third module expanded students' corpus research and querying skills. A combination of the top-down and bottom up approach was adopted, whereby students were guided to examining data retrievable from frequency lists, concordances of high-frequency words, and n-grams of function words with the support of the AntConc software. Worksheets with various tasks were designed to help them formulate search queries and focus on various typical features of specific domains and genres (See also Appendix 2). To discuss subject-specific terminology, students were also familiarized with the tools provided by the Sketch Engine platform (BootCat, Keywords, Word Sketches), which helped them extract key terminology and compile examples of frequent collocations. As recommended by Hubbard (2013), the results of their analyses were summarized in a collaborative debriefing at the end of each session, allowing students to compare their findings, become aware of linguistic variety within genres, registers and domains, as well as share procedures followed and challenges encountered during corpus consultation.

In the fourth module, corpus use was contextualised by demonstrating approaches and techniques that are relevant for language professionals, such as disambiguating language use, identifying subject-specific terminology and collocations. Students were asked to produce their own texts in English and translate short excerpts related to various specialist domains tackled in their DIY corpora by querying their corpora and available online corpora independently. The final sessions were dedicated to student presentations of the procedures and findings of their corpus-based analyses and evaluating their learning outcomes. In addition, students were asked to write a report discussing their corpus analysis and reflecting on their experience of autonomous corpus use.

Data collection and analysis

The study adopted a mixed-methods approach to examine students' perception of corpus use and their corpus consultation of online available and DIY corpora. It used the following main data sources: 1) student written reports and reflection on corpus consultation submitted at the end of the course; 2) an immediate post-course online questionnaire to measure students' perception of corpus use; and 3) an online questionnaire administered five months after the end of the course to establish autonomous corpus use.

Students were asked to write a written report and reflection on their corpus analysis and perceived learning achievements at the end of the course. Instructions for the report required students to compile a corpus of a minimum of 25,000 words and to analyse a minimum of 10 lexical items with a tool they used in the course. The analysis of the reports focused on three aspects: 1) the type of corpus resources, tools and techniques students used, 2) the depth of analysis and results presented, and 3) student reflection on corpus consultation. Corpus analysis techniques were applied to examine the third aspect of student reports. The reflection and conclusion sections of the student assignments were compiled into a corpus and a keyword analysis was carried out with the help of Sketch Engine. Concordances of keywords were analysed qualitatively for emerging categories of students' perceptions of corpus use (Charles, 2017).

The items in the questionnaires were created based on Yoon and Hirvela's (2004) and Charles' (2014) questionnaires. The immediate post-course questionnaire focused on students' perceived ease of use, perceived usefulness and general attitudes towards corpus use. To measure student's perception of corpus use 29 items were included. Students were asked to indicate their degree of agreement on a scale of 1 - 6 (1: strongly disagree, 2: disagree, 3: somewhat disagree, 4: somewhat agree, 5: agree, 6: strongly agree). For the analysis of Likert-scale items the mean scores were calculated. To enhance presentation of the data, responses were coded into two categories of 'agree' and 'disagree' by grouping all degrees of agreement into the 'agree' category and all degrees of disagreement into the 'disagree' category.

The second questionnaire was administered five months after the course and focused on the use of online available corpora and DIY corpora after the course. More recently studies on students' corpus use have applied computer-tracking logs (Pérez-Paredes, Sánchez-Tornel, Alcaraz Calero & Jiménez, 2011; Crosthwaite, Wong, & Cheung, 2019). The aim of these studies was to investigate students' engagement with corpora by analysing students' queries and interaction with corpus tools on university platforms. Using tracking software on students' own computers, which would have been the case in this study, would have raised a number of privacy and ethical issues and was likely to result in students' non-participation. Therefore, cognisant of the limitations of self-reporting, a questionnaire was considered appropriate for the purposes of the present study. The present study aimed to explore students' corpus use to establish whether participants had used corpora after the course and how useful they found corpus use. The questionnaire comprised 7 items. These included questions such as 'Have you used your DIY corpus at any time since the course ended?', 'Have you created another DIY corpus since the course ended?' and 'Have you used an available corpus at any time since the course ended?'. Likert scale questions were used to establish how useful and how easy or difficult students

found their autonomous corpus use for purposes outside the course requirements, for example, writing assignments or translations for academic or professional purposes (see Appendix 4).

Results and discussion

Analysis of written reports

As can be seen in Table 1, students chose a range of topics for their small DIY corpora, some focusing narrowly on a specialized area, others choosing a large field of knowledge. The sizes of their DIY corpora ranged from 25,000 to 100,000, which is comparable to DIY corpora created by students in previous studies (Charles, 2014; Jablonkai & Čebroň, 2017; Leńko-Szymańska, 2017). Three students used WebBootCat to expand their DIY corpora. The texts for the small DIY corpora were retrieved from Internet sources, for example, online news portals (e.g. The Guardian, ABC news), reports from international organisation websites (e.g. UN, EU) and research articles. Although students were instructed to either compile documents of only one genre (e.g. academic articles, transcripts of news programmes, news articles, official documents and reports), or else balance the word tokens of each genre included in the corpus, most students had difficulties retrieving the minimum required number of 25,000 words of the same genre. Thus, only two students compiled academic articles only, while seven reported encountering problems since most of the articles that surfaced in their searches for academic texts were only available in full text after purchasing them. Therefore, they proceeded to gather documents of various genres. Having had no previous experience of dealing with specialized language varieties, undergraduate students found it difficult to identify specialist domains or specific areas to study and explore. Students also observed that it took considerable time to compile an appropriate corpus as the following quote exemplifies: *“the process of compiling this corpus took a while as I was looking for diversity and relevance of texts I wanted to use for the research”* (S1). The reports also demonstrated that students were critically engaged with the corpus compilation process and considered carefully which texts to include: *“The most important step was to choose the right articles, but I selected articles from selected literature for the Course on the Media”* (S9). Furthermore, corpus compilation provided them with deeper insights into what constitutes corpus data and how to interpret corpus output. For example, one student pointed out that in her corpus *“the documents were all written in American English and represent usage in the American variety of English only.”* (S6). The reports in this study suggest that the direct involvement in the corpus compilation process had an added value for undergraduate students as was reported in previous studies with doctoral students (Charles, 2012). Their lack of experience with subject-specific language varieties, however, seems to necessitate more guidance in text selection, balancing and structuring their DIY corpora.

Table 1 DIY corpora compiled by students

Students	Topic	Genre	Word Tokens	WebBootCat
S1	Culture	Mixed genre	30,421	
S2	Climate change	Mixed genre	26,500	
S3	Education	Mixed genre	45,082	344,000
S4	Employment	Mixed genre	39,082	280,000
S5	Environment	Mixed genre	30,000	
S6	Food safety	Academic articles	26,000	
S7	Human Rights	Mixed genre	25,000	
S8	Linguistics	Mixed genre	25,000	140,000
S9	Media	Academic articles	102,365	

As regards the corpus tools, all students initially used the AntConc software to analyse their corpus. They therefore needed to convert all the PDF and Word files into Plain Text files and clean them of tables, footnotes and similar sections. Students found the use of AntConc very useful for analysing collocations and frequency lists, but its use was perceived as more challenging than using Sketch Engine: *“I used AntConc to study the concordances in terms of collocation patterns and bundles of most frequent lexical words. I found the programme very useful for its visual presentation of patterns, but more difficult to use than Sketch Engine”*. (S5)

All students found the Sketch Engine platform very user-friendly. They each compiled their own corpus on the platform, extracted the keywords and multi-word units by comparing their subject-specific corpora to a reference corpus (e.g. BNC, enTenTen15) and explored the Word Sketches of the words in the keyword list. They discovered that *“it depends on the chosen domain what the results will be in terms of keywords and their collocates”* (S3). Some students also expanded their corpora with the WebBootCat tool and compared the results obtained by both versions to find out that *“the expanded corpora were more similar to the general language, not so focused on the terminology of my specific domain”* (S6).

As regards the types and depth of analysis students carried out on their DIY corpora, the analysis revealed that students started by looking at frequency lists: *“Although the most frequent tokens were, as expected, mostly function words with no lexical meaning, such as prepositions, articles, auxiliary verbs, etc., I still sorted them in a list according to their frequency.”* (S6). Word frequency was also the most important criterion in selecting the words for initial exploration of the corpora. All students compiled the results of their research in the form of a list of most frequent collocations and n-grams of the ten most frequent keywords. This was presented in the format of a simple grid provided by the instructor.

Two students engaged in searches beyond the strict requirements given in the instructions. One student took the research a step further and reported on comparing the results of her research to information found in on-line dictionaries, thus discovering: *“The first thing I*

would like to point out is the importance of relevance of chosen texts. At first, I chose some random texts and I did not get interesting results, because the terms and collocations were the same as in a general dictionary. I had to search carefully and go through every text in detail in order to identify more relevant texts for my specific domain". (S8)

Another student (S3) sorted the first 100 lexical words from the frequency list according to their part of speech and commented on the extent to which each word can be considered an element of subject-specific terminology. She also exemplified her findings with concordances and compared the collocations yielded from her DIY corpus with those found in the enTenTen15 corpus. Her comments showed enhanced language awareness and an extended understanding of how corpora can represent language use: "it is also clear that a corpus best represents the language when it is compiled with up to date texts because every language is constantly evolving and new words and usages are emerging all the time." (S3). She also demonstrated a good grasp of the subject-specific nature of her specialised DIY corpus: "I learned that general corpora report different information from the specialized ones" (S3).

In order to investigate students' reflections on the benefits and challenges of their corpus use, a small corpus (2587 running words) of the reflection and conclusion sections of the written reports was compiled. Using the Keyword tool on the SketchEngine platform, 42 lemmas of lexical words (Biber, Johansson, Leech, Conrad and Finegan, 1999) that appeared at least 6 times in the corpus, were retrieved (See Appendix 3).

The following categories emerged from the analysis of the concordance lines of the identified keywords that are marked in bold in the quotes below.

1) Advantages of using a language corpus

Concordances of the following lemmas revealed the advantages students discussed in their reports: *corpus, language, vocabulary, learn, improve, study, use, student, academic, tool, important, even, very*. All student reports (100%) discussed corpus consultation as an important approach to acquiring languages in the academic environment and beyond. They emphasize that "a **corpus** is an indispensable **tool** in teaching and **learning** a foreign **language**, especially for expanding the **vocabulary** of the learners" (S7).

2) Affordances of corpus tools and resources

Specific affordances of corpus resources and tools were described with the following lemmas: *result, provide, see, useful, new, other, list, frequency, adjective, noun, verb, vocabulary, collocation*. These statements point to the new operations students learned to perform. All students (100%) mentioned the benefits of enhancing their vocabulary by corpus exploration, highlighting that "the Keyword **tool** helped me to create the **list** of most **useful** expressions" (S8) and "**Learning vocabulary** [...] through **corpora** [...] you can **see** straight away with which **adjective, noun and verb a word** is most **frequently used** with [...] and **frequency** shows which are weak or strong **collocations**" (S9).

3) Enhanced language awareness

Concordances of keywords such as *text, compare, research, find, know, different, more*, revealed enhanced language awareness and a clearer understanding of differences in varieties

of English. Around 91% of students mentioned these issues claiming that corpus research helped them discover “*differences between BE and AE and Australian English*” (S3) and that “*spoken and written language are not the same*” (S1), as well as find out “*that the term media can have various meanings in different specialist fields and in general language*” (S9).

4) Understanding the limitations of DIY corpora

However, students also realized the limitations and problems of corpus-based investigation of language. 44% of students pointed out such disadvantages by using keywords such as *important, text, collocation, corpus* to express these concepts. Some point out that “*we should be careful when collecting a corpus*” and “*select relevant texts*” (S2), but also that “*a too small corpus [will not] give accurate information*” (S6).

5) Future plans to use language corpora

While all the students expressed their positive perception of corpus consultation, only 33% of them hinted at their future plans to engage in autonomous corpus use. The keywords used to express such plans were *academic, corpus, vocabulary, language, new, study* and *use*. One student planned that “*In the future I would like to focus on compiling an academic corpus, [...]in order to expand my academic vocabulary and [...] as an aid for translation studies*” (S8). Another claimed that “*studying a corpus or compiling a new one will move us step by step to a better knowledge and use of language*” (S5).

The analysis of the written reports indicates that undergraduate students developed basic technical and corpus linguistic skills and used a few specific tools and functions of corpus analysis software confidently and appropriately. However, there were only two students who explored corpora and corpus tool functions beyond the explicit requirements of the assignments. The comments suggest that all students acknowledged the usefulness of a corpus-based ESP course and appreciated the benefits of corpus compilation and analysis. Students recognized the importance of focusing on ESP terminology, but did not seem to discern any particular variation in features of specific genres. These findings seem to indicate that corpus-based courses for undergraduate students need to pay more attention to variation in lexico-grammatical features in different genres.

Results of the two questionnaires

Overall, the immediate post-course questionnaire results show a similar positive picture when it comes to the perceived usefulness of corpus use. As can be seen in Table 2, mean scores in general are relatively high ranging from 3.75 to 5.33 on a six-point scale. This suggests that the majority of students in this undergraduate group found corpora helpful, they felt confident using a corpus, felt that corpus use made them more confident about their English learning and intended to use corpora in the future. Students found corpora most useful for learning the usage of words and phrases, which recorded the highest mean scores of 5.22 and 5.33 respectively. Results show mixed feelings when comparing corpora to Google search. It seems that students did not necessarily find corpus consultation more useful than using Google as half of the group agrees and the other half disagrees with this statement. Reasons

here might be that undergraduate students have different expectations from a language learning tool. Studies with doctoral students found that they also resorted to Google when they write because it was quicker and more convenient. However, they also reported that corpus use resulted in more accurate and specific output (Charles, 2014; 2017). Therefore, undergraduate corpus-based courses could include an explicit demonstration of how corpus consultation compares to Google searches and how corpora can provide more appropriate subject-specific lexico-grammatical patterns of usage. Students seem to see value in corpus use especially for writing (4.67), learning vocabulary in general (4.78) and for specific subject fields in the future (4.67), however, not so much for writing for other courses in English.

Table 2 Overall perception of corpus use

Question	Mean ^a	Agree	Disagree
1 Corpus use was easy to learn	4.67	9 (100%)	0 (0%)
2 The practice sessions in the computer lab were helpful for learning corpus use	4.22	7 (78%)	2 (22%)
3 I feel confident using a corpus.	4.89	8 (89%)	1 (11%)
4 A corpus is more helpful than a dictionary for learning English	4.11	8 (89%)	1 (11%)
5 A corpus is more useful than Google for learning English	3.75	4 (44%)	4 (44%)
6 Using a corpus is helpful for learning the meaning of words	4.67	8 (89%)	1 (11%)
7 Using a corpus is helpful for learning the usage of words	5.22	9 (100%)	0 (0%)
8 Using a corpus is helpful for learning the usage of phrases	5.33	9 (100%)	0 (0%)
9 Using a corpus is helpful for learning grammar	4.22	8 (89%)	1 (11%)
10 I want to use a corpus in my next English courses too	4.11	7 (78%)	2 (22%)
11 When I have problems in English writing, I search for help in a corpus	4.33	7 (78%)	2 (22%)
12 When I search for information in a corpus, I usually get the information that I need	4.44	8 (89%)	1 (11%)
13 I use a corpus when writing papers in English for other courses too	4	5 (56%)	4 (44%)
14 As I have learned more about corpora, I have come to like them more	4.22	6 (67%)	3 (33%)
15 I will use a corpus for my English writing in the future	4.67	7 (78%)	2 (22%)
16 I will use a corpus for learning English words in the future	4.78	8 (89%)	1 (11%)
17 I will use a corpus for learning English words of a specific subject-field in the future	4.67	8 (89%)	1 (11%)
18 Learning about corpora has increased my confidence about learning English	4.44	7 (78%)	2 (22%)
19 I will recommend corpus use to other students to improve their English	4.89	7 (78%)	2 (22%)
20 Corpus use should be introduced in all English language courses	4.56	7 (78%)	2 (22%)

a 1: strongly disagree 2: disagree 3: somewhat disagree 4: somewhat agree 5: agree 6: strongly agree

As regards the challenges with corpus use, mean scores range from 3.33 to 4.11 suggesting that students found corpora neither very difficult nor very easy to use. As can be seen in Table 3, the biggest challenge seems to be unfamiliar words in concordances and collocation output.

Interestingly, students did not seem to be overwhelmed by the high number of sentences in concordances, but rather they found it difficult to deal with the limited number of sentences in concordances. The reason for this might be the use of small DIY corpora that in some cases might not have included many occurrences of specific vocabulary items.

Table 3 Difficulties with corpus use

Question	Mean ^a	Agree	Disagree
21 I have some difficulty in using a corpus due to time and effort spent on analyzing the data	3.89	7 (78%)	2 (22%)
22 I have some difficulty in using a corpus due to unfamiliar words on concordance/collocation output	4.11	7 (78%)	2 (22%)
23 I have some difficulty in using a corpus due to cut-off sentences in concordance output	3.67	6 (67%)	3 (33%)
24 I have some difficulty in using a corpus due to too many sentences in concordance output	3.56	4 (44%)	5 (56%)
25 I have some difficulty in using a corpus due to the limited number of sentences in concordance output	3.67	7 (78%)	2 (22%)
26 I have some difficulty in analyzing concordance output	3.89	7 (78%)	2 (22%)
27 I have some difficulty in analyzing collocation output	3.67	5 (56%)	4 (44%)
28 I have some difficulty in performing corpus search	3.33	4 (44%)	5 (56%)
29 The real texts in a corpus are too difficult to understand	3.56	5 (56%)	4 (44%)

a 1: strongly disagree 2: disagree 3: somewhat disagree 4: somewhat agree 5: agree 6: strongly agree

Finally, results of the questionnaire that asked about corpus use after the course revealed that, except for one student, all students used online available corpora and they felt that they were fairly easy to use, and their use improved their English. As regards the use of their own DIY corpora, a third of the students reported using their DIY corpora and two of them reported to have created an additional DIY corpus after the course. Reasons for the low take-up of DIY corpora might partly have been that DIY corpora were specifically designed for students' projects for the course and had little value for other purposes.

Overall, findings suggest that undergraduate students found corpus use beneficial for their ESP learning. Students were likely to use corpora for learning words or for writing purposes in the future. Their reports reveal that they gained an improved understanding of subject-specific lexical and lexico-grammatical features and an enhanced awareness of different regional and subject-specific varieties of English. In general, students perceived corpus use as neither very easy nor very difficult. The perceived challenges included the time and effort involved in compiling their own DIY corpora, especially, selecting appropriate texts and balancing their DIY corpora in terms of genres. A majority found analysing concordance output challenging and the reasons were often the limited number of sentences and unfamiliar words. Contrary to findings of previous studies (Chen & Flowerdew, 2018a), students did not mention difficulties with technical issues, complexity and computer skills relating to the

corpus tools used. Students felt fairly confident about their acquired skills and were able to perform corpus consultation and interpret their results appropriately.

Conclusions

In this study we examined the outcomes and student perceptions of a corpus-based undergraduate ESP course. The course was designed following guiding principles that integrated relevant elements from DDL courses and learner training in CALL. It adopted a cyclical approach with repetitions and reflections as proposed by CALL training, introducing the elements of corpus analysis step-by-step and giving students enough time to gradually discover and internalise each new aspect of corpus analysis. The study also trialled the approach of teaching ESP with DIY corpora to undergraduate students. Findings suggest that students considered this approach useful. There are, however, some aspects that might need to be reconsidered based on student responses. First, the required minimum size of DIY corpora should be increased to at least 50,000 running words in order to yield relevant data. Second, the tasks for exploration of various lexico-grammatical aspects of language should also include comparisons based on online corpora in students' L1. This may further enhance students' engagement with corpus consultation after the course, helping them to internalise the corpus-based research principles more deeply. At the same time, such an approach may especially help enhance undergraduate students' language awareness. Participants in this study did not have a good understanding of the concept of collocation in general and did not seem to be aware of collocation patterns in their L1. Finally, DIY corpus compilation should be introduced at an earlier stage in the course as the compilation process provided students with the learning opportunity to evaluate and select texts for inclusion focusing on specialist domains. This critical engagement with corpus compilation seemed to help undergraduate students develop a deeper understanding of corpus structure, potential scopes of investigation and interpretation of output of corpora in general. Therefore, the introduction of online available corpora could be shortened and delivered at a later stage.

Although it is not possible to conclusively generalise from a small number of participants, findings seem to indicate that a DDL approach with DIY corpora has value for not only postgraduate EAP students, but also for professional ESP contexts at the undergraduate level.

References

- Ackerley, K. (2017). Effects of corpus-based instruction on phraseology in learner English. *Language Learning & Technology*, 21(3) 195-216. Retrieved from <http://llt.msu.edu/issues/october2017/ackerley.pdf>
- Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word combination In A. Cowie (Ed.) *Phraseology: Theory, analysis and applications* (pp. 101-121). Oxford: Oxford University Press.
- Anthony, L., 2014. AntConc (3.4.3). [computer program] Tokyo, Japan: Waseda University. Available at: <<http://www.laurenceanthony.net/>> [Accessed 29 August 2018].

- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson Education.
- Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, 67(2), 348–393. <https://doi.org/10.1111/lang.12224>
- Bridle, M. (2019). Learner use of a corpus as a reference tool in error correction: Factors influencing consultation and success. *Journal of English for Academic Purposes*, 37, 52–69. <https://doi.org/10.1016/J.JEAP.2018.11.003>
- Chambers, A., Farr, F., & O’Riordan, S. (2011). Language teachers with corpora in mind: From starting steps to walking tall. *Language Learning Journal*, 39(1), 85–104. <https://doi.org/10.1080/09571736.2010.520728>
- Charles, M. (2012). ‘Proper vocabulary and juicy collocations’: EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, 31(2), 93–102. <https://doi.org/10.1016/j.esp.2011.12.003>
- Charles, M. (2014). Getting the corpus habit: EAP students’ long-term use of personal corpora. *English for Specific Purposes*, 35(1), 30–40. <https://doi.org/10.1016/j.esp.2013.11.004>
- Charles, M. (2017). Do-it-yourself corpora in the classroom: Views of students and teachers. In L. Wong, & K. Hyland (Eds.), *Faces of English education: Students, teachers and pedagogy* (pp. 107–123). Abingdon: Routledge.
- Charles, M. (2018). Corpus-assisted editing for doctoral students: More than just concordancing. *Journal of English for Academic Purposes*, 36, 15–25. <https://doi.org/10.1016/j.jeap.2018.08.003>
- Chen, M., & Flowerdew, J. (2018a). A critical review of research and practice in data-driven learning (DDL) in the academic writing classroom. *International Journal of Corpus Linguistics*, 23(3), 335–369. <https://doi.org/10.1075/ijcl.16130.che>
- Chen, M., & Flowerdew, J. (2018b). Introducing data-driven learning to PhD students for research writing purposes: A territory-wide project in Hong Kong. *English for Specific Purposes*, 50, 97–112. <https://doi.org/10.1016/J.ESP.2017.11.004>
- Crosthwaite, P. & Cheung, L. (2019). *Learning the language of dentistry*. Amsterdam, Philadelphia: John Benjamins.
- Crosthwaite, P., Wong, L. & Cheung, J. (2019). Characterising postgraduate students' corpus query and usage patterns for disciplinary data-driven learning. *ReCALL*, 31(3), 255–275.
- Frankenberg-Garcia, A. (2005). Pedagogical uses of monolingual and parallel concordances. *ELT Journal*, 59(3), 189–198. <https://doi.org/10.1093/elt/cci038>
- Frankenberg-Garcia, A. (2012). Raising teachers’ awareness of corpora. *Language Teaching*, 45, 475–489. <https://doi.org/10.1017/S0261444810000480>

- Frankenberg-Garcia, A. (2018). Investigating the collocations available to EAP writers. *Journal of English for Academic Purposes*, 35, 93–104. <https://doi.org/10.1016/j.jeap.2018.07.003>
- Friginal, E. (2018). *Corpus linguistics for language teachers*. New York, London: Routledge.
- Gavioli, L. (2005). *Exploring corpora for ESP learning*. Amsterdam: John Benjamins.
- Godwin-Jones, R. (2017). Data-informed language learning. *Language Learning & Technology*, 21(3), 9–27. Retrieved from <http://llt.msu.edu/issues/october2017/emerging.pdf>
- Holec, H. (1981). *Autonomy in foreign language learning*. Oxford, England: Pergamon Press.
- Hubbard, P. (2013). Making a case for learner training in technology enhanced language learning environments. *CALICO Journal*, 30(2), 163-178.
- Hüttner, J., Smit, U., & Mehlmauer-Larcher, B. (2009). ESP teacher education at the interface of theory and practice: Introducing a model of mediated corpus-based genre analysis. *System*, 37(1), 99–109. <https://doi.org/10.1016/j.system.2008.06.003>
- Jablonkai, R. R., & Čebroň, N. (2017). Corpora as tools for self-driven learning. In N. A. Alias & J. E. Luaran (Eds.), *Student-driven learning strategies for the 21st century classroom* (pp. 274–298). Hershey, PA: IGI Global. <https://doi.org/10.4018/978-1-5225-1689-7.ch018>.
- Johns, T. (1991). Should you be persuaded: two examples of data-driven learning. In T. Johns & P. King (Eds.), *Classroom concordancing. ELR Journal*, 4 (Special Issue), 1-16.
- Johns, T. (2002). Data-driven learning: The perpetual challenge. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 107-117). Amsterdam: Rodopi.
- Lee, D. & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, 25(1), 56– 75.
- Leńko-Szymańska, A. (2014). Is this enough? A qualitative evaluation of the effectiveness of a teacher-training course on the use of corpora in language education. *ReCALL*, 26(2), 260–278. <https://doi.org/10.1017/S095834401400010X>
- Leńko-Szymańska, A. (2017). Training teachers in data-driven learning: Tackling the challenge. *Language Learning & Technology*, 21(3), 217–241. Retrieved from <http://llt.msu.edu/issues/october2017/lenko-szymanska.pdf>
- Long, M. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, R. Ginsberg, & C. Kramsch (Eds.) *Foreign language research in cross-cultural perspective*. Amsterdam: John Benjamins.

- O’Keeffe, A., & Farr, F. (2003). Using language corpora in initial teacher education: Pedagogic issues and practical applications. *TESOL Quarterly*, 37(3), 389–418. <https://doi.org/10.2307/3588397>
- Pérez-Paredes, P. (2019). A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. *Computer Assisted Language Learning*, DOI: 10.1080/09588221.2019.1667832.
- Pérez-Paredes, P., Sánchez-Tornel, M., Calero, J. M. A. & Jiménez, P. A. (2011). Tracking learners' actual uses of corpora: Guided vs non-guided corpus consultation. *Computer Assisted Language Learning*, 24(3), 233-253.
- Pinshuan, L., & Lin, H. (2019). The effect of the inductive and deductive data-driven learning (DDL) on vocabulary acquisition and retention. *System*, 81, 14–25. <https://doi.org/10.1016/j.system.2018.12.011>
- Sketch Engine Available at <http://www.sketchengine.eu>
- Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language and Communication*, 18(1), 47-67.
- Yoon, H. (2008). More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning & Technology*, 12(2), 31–48. Retrieved from <http://llt.msu.edu/vol12num2/yoona/>
- Yoon, H. & A. Hirvela (2004) ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, 13(4), 257-283.
- Yoon, H., & Jo, J. W. (2014). Direct and indirect access to corpora: An exploratory case study comparing students’ error correction and learning strategy use in L2 writing. *Language Learning & Technology*, 18(1), 96-117.

Appendix 1 Course outline

Corpus Topics	Lab practice
<i>Module 1: Core Concepts</i>	
Language corpora	on-line dictionaries vs corpus data
Collocations	on-line corpora in English and Slovene
Wordlists	frequency data, Academic Wordlist
Concordances	formulating queries
<i>Module 2: Corpus Compilation</i>	
DIY ESP corpora	purpose, type and size
Genres and registers	selecting documents
Harmonisation	corpus structure
Corpus consultation	pre-compiled corpora
<i>Module 3: Corpus Analysis</i>	
Corpus analysis tools	AntConc, SketchEngine
Corpus research	concordance manipulation
Analysing corpus output	word sketches, keyness
ESP terminology	BootCat: key words list, multi-word expressions
Corpus expansion	WebBootcat
<i>Module 4: Autonomous Corpus Use</i>	
Practicing corpus skills	Writing and translating
Presentations	Presentations of students' projects

Appendix 2 Sample lesson plan

AntConc tool: frequency lists, collocations & n-grams

AIMS:

- to identify the information that can be inferred from the frequency lists
- to extrapolate language patterns from a set of concordances
- to find elements of a word family

1. Introducing corpus analysis.
2. Using AntConc.
3. Analysing wordlists.
4. Manipulating concordances: frequent lexical words
5. N-grams and bundles

Left sort	Keyword	Right sort
	<i>Word no. 1</i>	
	<i>Word no. 2</i>	
	<i>Word no. 3</i>	
	<i>way</i>	

6. Identifying word families: wordlist - alphabetical order; use of wildcard*.
7. Comparing data from various ESP corpora: "accord", "evaluate", "way".
8. Discussing findings in group.

Appendix 3 Keywords in student reports

Raw frequency	Word	Frequency per 1000 words
46	corpus	17.8
36	use	13.9
29	text	11.2
29	word	11.2
22	adjective	8.5
20	verb	7.7
16	noun	6.2
15	vocabulary	5.8
14	language	5.4
12	research	4.6
10	frequent	3.9
10	study	3.9
10	learn	3.9
10	result	3.9
10	new	3.9
10	more	3.9
9	different	3.5
8	academic	3.1
8	very	3.1
8	other	3.1
7	useful	2.7
7	compare	2.7
7	list	2.7
7	important	2.7
7	provide	2.7
7	find	2.7
6	collocation	2.3
6	tool	2.3
6	see	2.3
6	know	2.3

Appendix 4 Questionnaire 2

1. Age:
2. Gender:
3. Nationality:
4. Study programme:
5. Did you take part in the Corpus-based ESP course in November? Y/N
6. Have you used your DIY corpus at any time since the course ended? Y/N
7. Has the use of your DIY corpus helped to improve your English?
Yes, to a great extent 1 2 3 4 5 6 No, not at all
8. Did you find creating your DIY corpus was
easy 1 2 3 4 5 6 difficult
9. Have you created another corpus since the course ended? Y/N
10. Have you used an available corpus at any time since the course ended? Y/N
11. Has the use of an available corpus helped you improve your English?
Yes, to a great extent 1 2 3 4 5 6 No, not at all
12. Did you find using available corpora was
easy 1 2 3 4 5 6 difficult