

Jana Kegalj, Mirjana Borucinsky

Faculty of Maritime Studies, University of Rijeka, Croatia
jana.kegalj@pfri.uniri.hr, mirjana.borucinsky@pfri.uniri.hr

Data-Driven Learning for Writing Skills Development

Abstract

Exposing students to corpus-informed research is a typical example of data-driven learning. This paper reports on the ways that corpora (i.e., text collections), corpus tools (i.e., software packages), and corpus methods (i.e., techniques for analysing corpus data) can be used to develop students' writing skills, while enabling them to improve their digital competencies, which is in line with current trends in education. The authors present and discuss the ways that corpus-derived materials can be developed for teaching writing skills with the goal of engaging students during their learning process and enabling them to conduct their own linguistic research. The data-driven learning (DDL) method based on corpus search was implemented in a specialized course aimed for doctoral students at the University of Rijeka Faculty of Maritime Studies. The course, implemented within the UNIRI CLASS A2 *Digital Citizenship—Innovations in Learning and Teaching* in 2022 project line, aimed to use the corpus-based data-driven learning method to develop students' academic writing skills. This was intended to make students more independent and autonomous in their learning, to enhance their digital skills, and to stimulate them to be more involved in their own learning.

Keywords: data-driven learning, corpora, academic writing, language resources

1. Introduction

The rapid progress and increasing availability of language technologies have greatly influenced various areas of language production and language teaching. Language technologies comprise various software and tools that include natural language processing, lexical computing, and speech technologies. They can be broadly divided into language resources, language tools, and commercial products. The most prominent among language resources are corpora, which enable the processing of large amounts of linguistic data. The application of corpora has had a strong influence on teaching and learning of foreign languages (cf. Campoy-Cubillo et al. 2010; Hunston 2022), and now there are many researchers who suggest ways to use corpora in the classroom or to create classroom materials, dictionaries, glossaries, or other useful resources. Corpora provide multiple ways of usage to address different learning needs, but in particular they enable students to become more autonomous

in their learning, to self-direct their learning, and to gain life-long learning skills.

First of all, users of corpora, both teachers and students (Kilfarriff and Kosem 2012), have to distinguish among different language technologies offered nowadays, namely language resources, language tools, and commercial products (Tadić 2003), to be able to understand what each of them has to offer. Language resources represent linguistic material that has been digitally systematized and can be used to perform various searches. They include corpora, linguistic collections, and digital dictionaries. Language tools are specialized programs, developed on the basis of language resources, which enable the processing of existing resources or the creation of new ones, such as Sketch Engine,¹ LancsBox,² AntConc,³ etc. Commercial products include dictionaries, spell checkers, grammar checkers, style checkers, machine translation tools, computer-assisted translation tools, etc., which offer linguistic checks for a fee.

Johns (1991) was one of the first to suggest that corpora can be used in the classroom as an effective way to engage students and make them active participants in the learning process, i.e., to allow them to discover language and language patterns. This was the basis for the DDL method, which uses large amounts of data as input for students to observe, analyse, interpret, explore, compare, hypothesize, and draw their own conclusions. Since then, corpora have been introduced in teacher education, translator training, teaching literature, and assessment (Flowerdew 2012). The use of corpora in teaching has been attested by various studies (cf. Kennedy and Miceli 2001; Kennedy and Miceli 2010; Cheng et al. 2003; Chambers and O'Sullivan 2004; Gaskell and Cobb 2004; Lee and Swales 2006; Boulton 2012; Chujo et al. 2012; Boulton and Cobb 2017; Vyatkina 2016, 2020). This type of teaching stimulates student motivation, develops critical thinking and lifelong learning skills. The role of the teacher changes significantly as the teacher becomes more of a guide, a mentor, an advisor. The approach of using corpora as a source of information is problem-based as students are confronted with a problem that they need to investigate. It is also a form of self-directed learning as students can take initiative, identify their problems and sources, and then apply the strategies they have acquired to solve the problem.

Having in mind all the contemporary challenges posed by the digital revolution and the need to adapt to new circumstances, the authors wanted to provide a platform for developing writing skills using available language technologies. To be more specific, at the beginning of their studies, students are expected to write a scientific paper and later their thesis. They frequently encounter problems with structuring their papers, gaining appropriate linguistic knowledge, and modifying their style of writing. The new course, conceived as an online course, was set up to

1 <https://www.sketchengine.eu/>

2 <http://corpora.lancs.ac.uk/lancsbox/>

3 <https://www.laurenceanthony.net/software/antconc/>

address this problem by introducing students to corpora and the data that can be extracted from them.

2. Applications of Corpora in Teaching

Corpora provide large amounts of authentic data in a particular form of output. Although they present a useful resource in education, corpora are still not widely implemented, especially at lower proficiency levels, particularly when it comes to direct application in class (cf. Vyatkina 2020; Hunston 2022).

They can be used directly by teachers and learners in class or indirectly in the production of textbooks and teaching materials (Römer 2011). Indirect applications of corpora include their use by researchers and educational material writers as resources for designing teaching syllabi, selecting the material for teaching in class, using word lists of key vocabulary in a particular register, using corpus data to design teaching materials or create textbooks (so-called corpus-informed materials). Direct applications of corpora in teaching refer to hands-on experience with corpora where teachers and learners search corpora in order to learn about language patterns, words, or phrases “in an autonomous way” (Bernardini 2002: 165). In class, students can work indirectly with corpus results in the form of filtered and printed concordances, or they can work directly with corpora, which has become possible owing to the availability of software and resources (e.g., AntConc, Skell,⁴ etc). Students’ work on corpora can also be classified as guided (also called direct learning) or unguided (referred to as data-driven learning).

The most commonly used corpus functionality in classrooms are concordances, that is, lines of immediate context around a searched keyword. Johns (1991) proposed using them as a way for students to infer meanings and study functions of words and thus become aware of some typical combinations. Other data and activities that can be used in teaching include various frequency lists and lists of collocations that can provide insight into word meaning, typical grammatical patterns, stylistic preferences, conceptual framework, and semantic network. Flowerdew (2009) also mentions that corpora can provide useful information on the behaviour of words, multi-word units, grammatical patterns, semantic prosody and semantic preference, as well as pragmatic and textual features. The author also emphasized that corpora are not only research tools but also pedagogic tools, as they raise students’ language awareness, engage their interest, and develop autonomy in learning.

Flowerdew (2012) also notes that despite all the advantages of using corpora in the classroom, direct applications of corpora have not been implemented that much. She attributes this to the problem of authentication, i.e., demonstrating the usefulness of corpora to students, and the problem of simplification, as unedited corpus data

4 <https://skell.sketchengine.eu/#home?lang=en>

might be too demanding for students. This is corroborated by Boulton (2017) who notes that DDL has an “impressive pedigree” but remains on the margins in practice.

Ädel (2010) states that little attention has been paid to the possibilities of implementing corpora in teaching writing skills, apart from teaching vocabulary and collocations. She also emphasizes that the few examples of hands-on application of corpora in teaching writing were mainly one-time experiments. Farr and Karlsen (2023) note that DDL has mostly been introduced in higher education, with academic writing and teacher education as the most prominent areas. Here we will present a selection of case studies in which DDL was implemented to develop writing skills. The courses that used corpora directly in class focused mainly either on rhetorical functions (to gain fluency and wider knowledge about a specific genre or linguistic functions) or on specific lexical or grammatical structures (for accuracy and error correction).

2.1 DDL Focusing on Lexical and Grammatical Items in Writing

Focusing on teaching vocabulary, Thurstun and Candlin (1998) developed corpus-based learning materials that would cater for the needs of students from different disciplines, following the hypothesis that students generally do not struggle with discipline-specific vocabulary but have problems with mid-frequency vocabulary (cf. Li and Pemberton 1994; Coxhead 2000), or so-called “semi-technical” or “academic vocabulary” (Nation 1990). Thurnstun and Candlin (1998) focused on a restricted set of vocabulary items and the use of concordancing techniques to expose students to the authentic use of these items. They used key words from identified rhetorical functions (e.g., stating the topic, referring to the research literature, reporting the research of others, expressing opinions tentatively, explaining procedures in the study, drawing conclusions) to focus on specific vocabulary items.

Yoon and Hirvela (2004) implemented a corpus-based approach to learning writing in an intermediate and an advanced ESL class. The corpus approach was gradually introduced into the ESL course with the aim to enrich the content and teaching methodology. They took a gradual approach to corpus use, that is, from explanation and demonstration to the independent use of primarily concordances and frequencies. The use of corpus tools was aimed at enhancing students’ knowledge of vocabulary and grammar. The authors found that the level of student proficiency should be one of the key factors when considering how to incorporate corpus search into an ESL course.

Chambers and O’Sullivan (2004) introduced corpus consultation work into a master’s course to assist students in error correction of their own written work. A special part of the course focused on training students in corpus consultation skills. The lecturer provided feedback on a piece of writing produced by the students, and then the students had to correct and improve their text using corpus tools. Similarly, Gilmore (2009) intergrated corpus tools in the redrafting stage of students’

writing where they were required to make their own hypotheses on how to improve their writing based on corpus data.

Ädel (2010) also reports on a small-scale experiment conducted with beginner-level students aimed to improve their writing skills. The author used concordancing tools to target specific research questions, mostly in the form of guided learning. However, just like Lee and Swales (2006), Ädel's attempt did not become part of regular writing instruction.

Another attempt to introduce corpora in writing instruction was made by Kennedy and Miceli (2010). They used corpus work as an aid to writing, achieving accuracy, and solving a specific grammatical issues. As their students had an intermediate knowledge of English, they decided to apply a more guided approach to using corpora, calling it “corpus apprenticeship” and advocating for “corpus-consultancy literacy”, as the knowledge gained on such a course would be potentially used in future writing tasks. The authors aimed to teach students to use concordances to enrich the content and language and to edit their text.

Flowerdew (2012) also reported on a corpus-informed course aimed at writing reports. Students were gradually introduced to corpus consultation strategies during the course and not in separate sessions. The learning process was guided, with teacher-directed tasks until the last stage when students had to work with corpora on their own. The author adopted a “guided inductive approach” to assist students in interpreting concordance results as this posed a particular challenge to them. The author reports that the search queries mainly focused on lexical and grammatical elements, but they were also encouraged to observe other (e.g., phraseological or genre-based) properties.

Bruce et al. (2016) reported on conducting a series of workshops aimed at improving academic writing skills of chemistry students using the DDL approach. The authors compiled the FOCUS corpus of academic texts produced by students and regarded as strong examples of academic work by various departments at Durham University. Their workshops focused on academic voice, reporting verbs, nominalisation, punctuation, and connectives, all of which were analysed using examples from the corpus.

2.2 DDL Focusing on Rhetorical Functions in Writing

Bernardini (2004) described an approach she calls “corpus-aided discovery learning” which relied on autonomous learning with corpora as resource. In Bernardini's unguided approach, the learner is seen as a researcher who makes hypotheses, poses questions, and finds ways to combine corpus tools to solve them. The author also distinguishes between “learning *from* corpora” and “learning *with* corpora”, the latter being based on discovery learning. In the course, students worked with the instructor as their guide and explored language items and patterns in context to make inferences about their function in the text. Tribble (2004) uses a corpus

of business correspondence to show how keyword analysis can be an efficient tool in understanding linguistic patterns in professional surroundings. The keywords were then related to social functions in the texts. Lee and Swales (2006) reported on a corpus-informed writing course English for Academic Purposes (EAP) for doctoral students which relied mostly on concordancing and was focused on discourse features of academic texts. During the course, students consulted specialized corpora of academic writing and had to compile two additional corpora. Their final report had to focus on how corpus linguistics techniques raised their level of rhetorical consciousness. Cortes (2007) reported on a corpus-based genre-oriented EAP course aimed at raising students' awareness of different linguistic features typical of a genre and stimulating them to use this knowledge in the future.

Charles (2007, 2011) also described a writing course offered at the Oxford University Language Centre based on rhetorical functions in which concordancing was used to focus on specific lexicogrammatical units in which a particular function was realized. The corpus used consisted of theses written by native speakers. In this course, the author combines the top-down discourse-based approach with the bottom-up corpus-based approach to make students aware of the key features of academic discourse. Hyland (2012) similarly emphasized the usefulness of corpora in identifying lexicogrammatical regularities in academic texts.

Flowerdew (2015) designed a workshop for science and engineering students in which corpus tools were used to identify typical move structures in the discussion section of a thesis. The students were familiarized with search strategies and then they identified useful lexical and grammatical items for specific rhetorical functions.

Aull (2017) conducted a course using context-informed corpus analysis for first-year college students with the aim of identifying the key structures used in the macro-genres of argumentative and explanatory writing. Chen and Flowerdew (2018) held a series of DDL workshops for PhD students focusing on discourse features of EAP and the language structures used to perform specific functions in a scientific text. Similarly, Dong and Lu (2020) reported on a corpus-based and genre-based EAP course set out to teach rhetorical structures in a discipline-specific academic writing course. Similarly, Wong (2019) reports on postgraduate writing courses which use a corpus-based multidisciplinary thesis writing support resource developed for this purpose.

All of these examples (which are by no means exhaustive) demonstrate how prolific the area of DDL implementation in teaching writing actually is. The researchers also report an increase in their students' proficiency, autonomy in learning, and greater motivation, which was also the goal of the course described here.

2.3. Methodology of Corpus-Based Teaching of Writing Skills

In the course designed as part of the UNIRI CLASS A2 *Digital Citizenship—Innovations in Learning and Teaching* project, the students work directly with corpora. As

different studies on the application of corpora in class have shown (cf. Thurnston and Candlin 1998; Gaskell and Cobb 2004; Kennedy and Miceli 2010), learners benefit more when the course progresses from more indirect, guided tasks towards direct, unguided tasks, which was adopted in this course as well, with the aim of reducing the students' dependence on the instructor's help. In this way, it is assumed that the students will gain basic knowledge on corpora and then have more control of their learning and become more independent. The goal was to demonstrate the usefulness of corpus tools in everyday writing tasks of PhD students, stimulate them to further explore different linguistic variations, promote autonomy in online surroundings, and use corpus tools to improve their own previously written works. Besides that, such learning ultimately promotes "tailor-made" learning, as it does not determine exactly what should be learnt, thus stimulating what Bernardini calls "serendipitous learning".

The design and topics included in the course were based on corpus studies of academic writing, which has significantly contributed to our understanding of this register. Specifically, corpus research has singled out a range of lexical and grammatical features typical for academic discourse, as well as described rhetorical functions used in different genres. As the course is intended for PhD students from different fields, the examples were taken from academic texts from different areas, while the focus was on general features of academic texts and so-called academic vocabulary which is common to all disciplines. The structure of the course adhered to Charles's (2011) proposed three-stage process: first, corpus awareness to introduce students to corpora and the data that can be extracted from them; second, corpus literacy to make the students perform simple searches, to understand concordance data, and to be able to make their own queries; third, corpus proficiency to make the students build their own (however small) corpus and perform queries in that corpus.

In the first, guided part of the course, the students work on a ready-made corpus of academic writing compiled in the AntConc program (Anthony 2023).⁵ The students are introduced to the AntConc program through demonstration of its features and opportunities to use the program directly. After that, they are assigned various guided tasks with frequency lists, concordances, collocation lists from other programs (e.g., Sketch Engine, LancsBox) to familiarize them with other applications and different display options that these programs offer. In the second, unguided part of the course, the students are encouraged to compile a corpus by themselves, collecting texts from their area of interest. After compiling the corpus, they are tasked with writing a journal article abstract, which is divided into several steps. Thurnston and Candlin's chain of activities (1998: 272), namely "look at concordances", "familiarize yourself with the patterns", "practice key terms", and "create your own writing", were modified here into more research-directed activities:

⁵ The corpus used was AmE06_Learned, consisting of 80 files, a total of 161469 tokens.

analyse your data, deduce general rules, produce your work based on previous findings. The tasks were designed to combine lexical and grammatical features with discourse functions they realize in the text.

3. A Sample of Practical DDL-Based Tasks Employing Corpora

One module of the course focuses on writing journal abstracts, which is a challenging form, as it needs to provide the reader with a short summary of the research topic, methodology, and results. The goal was to enable students to write an abstract for their own written article by analysing the lexical and grammatical items used for a particular purpose in order to raise their awareness of the rhetorical functions found in these short texts. The students worked on the corpus of abstracts that they had compiled. The course was not designed to be comprehensive and all-inclusive, but the primary objective was to develop the students' DDL skills so that they themselves can continue using this approach for their future learning needs.

The module starts with a series of guided tasks to stimulate the students to use corpus tools to distinguish between words “paper”, “article”, “work”, “study”, and “research”. In order to solve the tasks, the students need to use different functionalities of corpus processing tools (e.g., concordances, frequency analysis, list of collocates, n-grams). Some research (cf. Thurnstun and Candlin 1998) has shown that over-exposure to concordance lines might be tiring to students. Hence, a variety of tasks was designed ranging from closed exercises such as the “one-item multiple contexts” exercise (Johns 2000) shown in Figure 1 to open tasks such as the one shown in Figure 2.

<p>Tactics, an article discussing the political direction Russia should take after the revolution. In this conceptualisation. The final chapter will be devoted to a direct critique of Cruickshank's recent article. Canaan Banana, the former president of Zimbabwe, is quoted by Michael Edwards in his 1989 article, London/New York, 1999) Wyatt, Mark, White Knuckle Ride. (Salamander Books, London, 1996) as a scientific method of observation. James presented his theories by writing books and journal articles. The Levellers set out their franchise proposals to be debated at Putney in 1649. On the latter point Walwyn responded unequivocally, referring to 'in jedem Regiment' - here we see a slight indication that, although the press prevents the publication of articles - a philosophy which went against what the Ancien Régime represented. The first required shift in attitude was to take place. This can be seen in the language used for the eleventh article and that it merely 'took the form of a short, concise preamble to the constitution'. Indeed, on des droits de l'homme was, in some ways, a machine de guerre contre l'Ancien Régime. The Cambridge Law Journal, vol. 60, no. 2, Human Rights Act 1998 s.6. R (S) v Chief Constable of South Yorkshire, this declaration of incompatibility is sought with regard to open question. However, retention of photographs and fingerprints were seen not to contravene Article 8(1) of the ECHR. Lord Steyn then underscores that 'disclosure of private information by State Institutions' as an interference with Article 8(1) of ECHR. She fears that if the model is also erroneous and is in need of further refinement. There is no doubt that this is one of the most controversial of the four, taking what Brown said out of context and too the extreme (column and Discuss How the Results of This Article Support or Challenge This Developmental Theory. The d</p>	<p>he advocated a 'revolutionary democratic dictatorship of the proletariat and peasantry', a term that Marx A tale of two ontologies: an immanent critique of critical realism (The Sociological Review (54), 2004). entitled, "The irrelevance of development studies" as follows, "Whereas an armchair intellectual of r. Bennett, Peter, Blackpool Pleasure Beach, Burk, John, Tour of Blackpool, Auge, M, Orienting, Chapter 1 , teaching them as a lecturer, and giving a series of 'Talks to Teachers'. He is known for trying to I from An Agreement of the People, the first Agreement, which stated "That the people of England, t XXX of the final Agreement. Morton, The World of the Ranters, p. 183. That it shall not be in against the state's wishes, it cannot stop the writing or thinking of such ideas and this point offers a slight of the Déclaration states that: "Les hommes naissent et demeurent libres et égaux en droits." and of the Déclaration, whereas others are a list of rights, this is almost a call to reason, hinting that without 16 in the document itself refers to a constitution directly hinting that this document does not have the por themselves attempt to destroy all that was unjust in French Society at that time concerning national sov 2, Protocol I, European Convention on Human Rights. De Freitas v Permanent Secretary, Min 64(1A) of the Police and Criminal Evidence Act 1984 (PACE) which allows Police to retain DNA samples. 8(1) of ECHR in Kinnunen v. Finland in relation to fraud. Therefore, according to Lord Ste 14 is 'not limitless' and going beyond the prescribed boundaries is not its goal. If this is discrimination, A 8(1) is not engaged in this case then neither is Article 14 and therefore the State becomes free to retain is the most controversial of the four, taking what Brown said out of context and too the extreme (column and Discuss How the Results of This Article Support or Challenge This Developmental Theory. The d 170) Treaty of Amsterdam (EC Treaty) Article 228 EC (ex. Article 171) Treaty of Amsterdam (EC Treaty).</p>
--	--

Figure 1. Example of a one-item multiple contexts exercise.

In this particular task, the students are asked to write down the different meanings of the word “article” that can be extracted from the concordance lines (the

concordance lines were taken from Sketch Engine). Different tasks target various problem-solving techniques by the student.

Tactics, an article discussing the political direction Russia should take after the revolution. In this article he advocated a 'revolutionary democratic dictatorship of the proletariat and peasantry', a term that Marx conceptualisations. The final chapter will be devoted to a direct critique of Cruickshank's recent article A tale of two ontologies: an immanent critique of critical realism (The Sociological Review (54), 2004). Caneana Banana, the former president of Zimbabwe, is quoted by Michael Edwards in his 1989 article entitled, "The irrelevance of development studies" as follows: "Whereas an armchair intellectual of ruralledge, London/New York, 1999) Wyatt, Mark, White Knuckle Ride (Salamander Books, London, 1996) Articles Bennett, Peter, Blackpool Pleasure Beach, Burk, John, Tour of Blackpool, Auge, M, Orienting, Chapter 1 as a scientific method of observation. James presented his theories by writing books and journal articles, teaching them as a lecturer, and giving a series of 'Talks to Teachers'. He is known for trying to crux of his argument. The Levellers set out their franchise proposals to be debated at Putney in Article I from An Agreement of the People, the first Agreement, which stated: "That the people of England, t such beliefs on the movement. On the latter point Walwyn responded unequivocally, referring to Article XXX of the final Agreement. Morton, The World of the Ranters, p. 183. "That it shall not be in n jedem Reglement" - here we see a slight indication that, although the press prevents the publication of articles against the state's wishes, it cannot stop the writing or thinking of such ideas and this point offers a slight ct those rights - a philosophy which went against what the Ancien Régime represented. The first article of the Déclaration states that: "Les hommes naissent et demeurent libres et égaux en droits." "and anquired shift in attitude was to take place. This can be seen in the language used for the eleventh article of the Déclaration, whereas others are a list of rights, this is almost a call to reason, hinting that without mplete' and that it merely 'took the form of a short, concise preamble to the constitution'. Indeed, article 16 in the document itself refers to a constitution directly hinting that this document does not have the po on des droits de l'homme was, in some ways, a machine de guerre contre l'Ancien Régime. The articles themselves attempt to destroy all that was unjust in French Society at that time concerning: national sov Cambridge Law Journal, vol. 60, no. 2. Human Rights Act 1998 s.6. 53 Article 2, Protocol I, European Convention on Human Rights. 54 De Freitas v Permanent Secretary, Min e, R (S) v Chief Constable of South Yorkshire, this declaration of incompatibility is sought with regard to Article 64(1A) of the Police and Criminal Evidence Act 1984 (PACE) which allows Police to retain DNA samples open question'. However, retention of photographs and fingerprints were seen not to contravene Article 8(1) of ECHR in Kinnunen v Finland in relation to fraud. Therefore, according to Lord Ste ment between Appellants and people in an 'identical situation'. Lord Steyn then underscores that Article 14 is 'not limitless' and going beyond the prescribed boundaries is not its goal. If this is discrimination, A of private information by State Institutions" as an interference with Article 8(1) of ECHR. She fears that if Article 8(1) is not engaged in this case then neither is Article 14 and therefore the State becomes free to retain all of the model is also erroneous and is in need of further refinement. There is no doubt that this article is the most controversial of the four, taking what Brown said out of context and too the extreme (column of the Main Developmental Theories in Your Discussion Review A Recent 'Child Development' Journal Article and Discuss How the Results of This Article Support or Challenge This Developmental Theory. The d OF LEGISLATION Article 226 EC (ex. Article 169) Treaty of Amsterdam (EC Treaty) Article 227 EC (ex. Article 170) Treaty of Amsterdam (EC Treaty) Article 228 EC (ex. Article 171) Treaty of Amsterdam (EC Treaty).

Figure 2. Example of an open task based on concordance lines.

The series of tasks in the following example focuses the students' attention on different features of the given words, aiming to clarify their use. However, the deeper goal is to demonstrate to the students through their own hands-on experience with DDL how they can solve a different problem in their future work.

Example 1. A sample of tasks related to the use of the following words: 'paper', 'article', 'work', 'study', 'research'.

1. Open AntConc and upload the corpus AmE06_j_learned, which is a corpus of academic texts (see the AntConc tutorial if you do not know how to do that). Type in the word 'paper' and 'article' and analyse their concordances, plots, collocations, and clusters. What can you see? What is the frequency of 'paper' vs the frequency of 'article'?
2. Take a closer look at concordance lines and write down some of the main findings regarding the use of 'paper' and the use of 'article' in context. What are the similarities/differences?
3. What is the most frequent word to the left/right of the words 'paper' and 'article'?
4. Now type the word 'work' in AntConc in the same corpus (AmE06_J_learned). Which meaning does the word 'work' NOT refer to?
 - a) labour
 - b) research
 - c) article
 - d) effort

5. Study vs research

These two nouns seem similar, but they actually refer to different things.

Take a look at this list of collocates.

What do you think—are they collocates of the word ‘study’ or the word ‘research’?

6. Try searching for the word ‘study’ and the word ‘research’ in AntConc and find out which prepositions follow them.

Which is the most frequent preposition after the word ‘research’?

7. Now take a closer look at the word ‘research’ in the corpus. Which verbs follow this noun?

8. Write a sentence using the word ‘research’ and one of the verbs you have found in the corpus. The sentence should refer to your own current research, the research that you are currently reading about, or the research that you have read about.

One of the steps of the module focuses on general features of abstracts. The students analyse the length of the abstracts and arrangement of rhetorical moves (e.g., background, present research, methods/materials, results, conclusion, recommendations, etc.) in the corpus that they have compiled on their own. This is followed by a series of tasks linking lexical and grammatical structures to these moves, e.g., they have to analyse the tenses used in abstracts and then link the use of tenses with particular rhetorical moves. Another example of a corpus-based task is the analysis of the use of “I” and “we” and their frequencies in their corpus, as well as of instances of self-referring (e.g., “this article”) to make the students aware of this feature so they can use it in their writing. It is important to emphasize that the students have previously been familiarized with the corpus query software and have performed basic searches.

Example 2. A sample of tasks aimed at studying the general features of abstracts using a tailor-made corpus.

1. Take a look at the abstracts that you have collected in the previous step.

Can you calculate how many words they have? How many sentences do they have? What is the average number of words per sentence?

2. Take a look at your abstracts again. Which tense is mostly used?

3. Now that you have found the tense that is mostly used in your abstracts—can you say why do you think this tense is used more than other tenses? Write your opinion.

4. Using your corpus of abstracts in AntConc, find out about the use of pronouns ‘I’ and ‘we’ in the corpus. Which tool would be best to find that out?

Which pronoun is used more, if any? Is perhaps the use of expressions like ‘present authors’ used instead?

5. Now take a look at another feature of abstracts, which is self-referring. This implies that there is reference to the article/study/paper itself in the text, e.g., 'this article'. Are there such references in your corpus? Which tool would you use to find this out?

6. Go back to File view in AntConc and read a few of the abstracts that you have collected. Can you find a pattern? Is there a sequence of information that the authors are providing? What do they present first: purpose, background, methods, some results, or findings?

Write the structure of your abstracts here.

7. Let's now take a closer look at the opening sentences in the abstracts. Go to AntConc again and click on File view.

Go through the first sentences of your abstracts. What do they state: purpose? some standard practice? present action of the researchers? problem?

Write down 2-3 most common functions of opening sentences in your corpus.

8. Now go back to the same abstracts. Focus on the second sentence in each abstract. How is it linked with the previous one?

Research suggests that there are several ways to do this:

1. keeping the same subject,
2. putting the information from the second half of the first sentence in the subject position at the beginning of the second,
3. using a new previously unmentioned topic.

Which one of these was most common in your corpus of abstracts?

9. Researchers have established that it is quite common to introduce a new topic in the second sentence because articles have a limited number of words, they are highly compressed texts, and the authors expect the readers to have the relevant knowledge, so they do not go into further detail.

Study your corpus in AntConc. Take a look at n-gram results and single out those that would be used to connect two sentences. Write them down here and specify their meaning (contrast, addition, cause-effect, sequence, conclusion, purpose, etc).

At the end of this module, the students are expected to produce an abstract of their own. The abstract is then peer-reviewed by other participants in the course and by the instructor as well. The goal is for the students to produce a text with the assistance of DDL, i.e., corpus tools, and thus gain life-long skills they can use in any future writing project.

4. Conclusion

In this case study, corpus data were used as input for guided writing tasks and as the basis for productive tasks. The case study also took into account that students need to be trained in corpus research before they can engage in autonomous

corpus-informed learning. The process of material design highlighted some possible pitfalls and challenges of such an approach, mainly related to the compilation of a suitable corpus, the selection of appropriate material, the time required for students to master the skills and query the corpus, and the method of training learners to use corpora. The main goal of the course is to provide students with useful new skills that can be applied in any kind of writing task.

The course aims to use two types of corpora: ready-made corpora of academic texts at the beginning of the course and tailor-made corpora compiled by students at the more advanced level of the course. Ready-made corpora of academic texts are used to engage students in studying the frequent expressions, structures, and style of academic texts. After gaining some knowledge about corpora and how they can be studied, students compile their own corpus on a topic of their interest. The process of upgrading their writing skills is research-based and task-based using the data from the corpus. Such teaching poses many challenges. First, students do not possess the meta linguistic knowledge about corpora or linguistics so this aspect can be time-consuming and students might get discouraged by such an approach. Second, the availability of corpus analysis tools and ready-made specialized corpora is restricted (cf. Anthony 2019). Third, there are few resources and materials developed for corpus-based teaching of writing. This paper aims to address these challenges and bridge the said gaps by demonstrating the kind of corpus-based tasks that can be designed for learners.

The future endeavours in this sense will be aimed at determining how students have benefited from this approach, how they perceived this way of learning, and what advantages they found in such learning. Also, as a follow-up after they finish the course, a further goal is to establish whether they continue using these tools in their academic writing.

References

- Ädel, Annelie. 2010. "Using Corpora to Teach Academic Writing: Challenges for the Direct Approach." In Campoy, M. C., Bellés-Fortuno, B. & Gea-Valor, M. L.. (Eds.) *Corpus-Based Approaches to English Language Teaching*. London and New York: Continuum International Publishing Group. 39–55.
- Anthony, Lawrence. 2023. AntConc (Version 4.2.4) [Computer Software]. Tokyo, Japan: Waseda University. <https://www.laurenceanthony.net/software>
- Anthony, Laurence. 2019. "Tools and Strategies for Data-Driven Learning (DDL) in the EAP Writing Classroom." In Hyland K. & Wong, L. L. C. (Eds.) *Specialised English: New Directions in ESP and EAP Research and Practice*. London and New York: Routledge. 179-194.
- Aull, Laura. 2017. "Corpus analysis of argumentative versus explanatory discourse in writing task genres." *Journal of Writing Analytics* 1. 1–45.
- Bernardini, Silvia. 2002. "Exploring new directions for discovery learning." In

- Kettemann, B. & Brill, G. (Eds.) *Teaching and learning by doing corpus analysis. Proceedings of the fourth international conference on teaching and language corpora*. 165–182.
- Bernardini, Silvia. 2004. “Corpora in the classroom: An overview and some reflections on future developments.” In Sinclair J. (Ed.) *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins. 15–36.
- Boulton, Alex. 2009. “Corpora for all? Learning styles and data-driven learning.” In Mahlberg, M., González-Díaz, V. & Smith C. (Eds.) *Proceedings of 5th Corpus Linguistics Conference*.
- Boulton, Alex. 2012. “Hands-on/hands-off: Alternative approaches to data-driven learning.” In Thomas, J. & Boulton, A. (Eds.) *Input, process, and product: Developments in teaching and language corpora*. Masaryk University Press. 152–168.
- Boulton, Alex; Cobb, Tom. 2017. “Corpus use in language learning: A meta-analysis.” *Language Learning* 67(2). 348–393. doi: 10.1111/lang.12224
- Bruce, Megan Laura; Coffey, Philippa; Rees, Simon; Robson, Jacquie. 2016. “Write on the edge: using a chemistry corpus to develop academic writing skills resources for undergraduate chemists.” *Chemistry Education Research and Practice* 17. 580–589. doi: 10.1039/c6rp00005c
- Campoy-Cubillo, Mari Carmen; Bellés-Fortuño, Begoña; Gea-Valor, Maria-Lluïsa. 2010. “Introduction to Corpus Linguistics and ELT.” In Campoy-Cubillo, M. C., Bellés-Fortuño, B. & Gea-Valor, M.-L. (Eds.) *Corpus-Based Approaches to English Language Teaching*. London and New York: Continuum International Publishing Group. 3–17.
- Chambers, Angela; O’Sullivan, Íde. 2004. “Corpus consultation and advanced learners’ writing skills in French.” *ReCALL* 16(1). 158–172.
- Charles, Maggie. 2007. “Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions.” *Journal of English for Academic Purposes* 6. 289–302.
- Charles, Maggie. 2011. “Using hands-on concordancing to teach rhetorical functions: evaluation and implications for EAP writing classes.” In Frankenberg-Garcia, A., Flowerdew, L. & Aston, G. (Eds.) *New Trends in Corpora and Language Learning*. London: Continuum. 26–43.
- Chen, Meilin; Flowerdew, John. 2018. “Introducing data-driven learning to PhD students for research writing purposes: A territory wide project.” *English for Specific Purposes* 50. 97–112.
- Cheng, Winnie; Warren, Martin; Xun-feng, Xu. 2003. “The language learner as language researcher: putting corpus linguistics on the timetable.” *System* 31. 173–186.
- Chujo, Kiyomi; Laurence, Anthony; Oghigian, Kathryn; Uchibori, Asako. 2012. “Paper-based, computer-based, and combined data-driven learning using a web-based concordance.” *Language Education in Asia* 3(2). 132–145.
- Cortes, Viviana. 2007. “Exploring genre and corpora in the English for academic

- writing class.” *The ORTESOL Journal* 25. 8–14.
- Coxhead, Averil. 2000. “The Academic Word List: A Corpus-based Word List for Academic Purposes.” In Kettemann, B. and Marko, G. (Eds.) *Proceedings of the Fourth International Conference on Teaching and Language Corpora*.
- Cresswell, Andy. 2004. “Getting to ‘know’ connectors? Evaluating data-driven learning in a writing skills course in Corpora in the Foreign Language Classroom.” In Hidalgo, E., Quereda, L. & Santana, J. (Eds.) *Selected papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6)*. Amsterdam and New York: Rodopi. 267–288.
- Farr, Fiona; Hagen Karlsen, Petter. 2023. “DDL Pedagogy, Participants, and Perspectives.” In Jablonkai, R. R. & Csomay, E. (Eds.) *The Routledge Handbook of Corpora and English Language Teaching and Learning*. London and New York: Taylor & Francis. 329–343.
- Flowerdew, John. 2009. “Corpora in Language Teaching.” In Long, M. H. & Dougherty, C. J. (Eds.) *The handbook of language teaching*. Wiley-Blackwell. 327–350.
- Flowerdew, Lynne. 2012. “Corpora in the Classroom: An Applied Linguistic Perspective.” In Hyland, K., Huat, C. M. & Handford, M. (Eds.) *Corpus Applications in Applied Linguistics*. Continuum International Publishing Group: London and New York. 208–224.
- Flowerdew, Lynne. 2015. “Using corpus-based research and online academic corpora to inform writing of the discussion section of a thesis.” *Journal of English for Academic Purposes* 20. 58–68. doi: <http://dx.doi.org/10.1016/j.jeap.2015.06.001>
- Gaskell, Delian; Cobb, Tom. 2004. “Can learners use concordance feedback for writing errors?”. *System* 32(3). 301–19.
- Gilmore, Alex. 2009. “Using online corpora to develop students’ writing skills.” *ELT Journal* 63(4). 363–372.
- Hyland, Ken. 2012. “Corpora and Academic Discourse.” In Hyland, K., Huat, C. M. & Handford, M. (Eds.) *Corpus Applications in Applied Linguistics*. Continuum International Publishing Group: London and New York. 30–46.
- Hunston, Susan. 2022. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Johns, Tim. 1991. “Should you be persuaded: Two samples of data-driven learning materials.” In Johns, T. & King, P. (Eds.) *Classroom concordancing: ELR Journal*, 4. Birmingham: Centre for English Language Studies, University of Birmingham. 1–16.
- Johns, Tim. 2000. “Data-driven Learning: The Perpetual Challenge.” In Kettemann, B. & Marko, G. (Eds.) *Proceedings of the Fourth International Conference on Teaching and Language Corpora*. Amsterdam and New York: Rodopi. 107–118.
- Kennedy, Claire; Miceli, Tiziana. 2001. “An evaluation of intermediate students’ approaches to corpus investigation.” *Language Learning & Technology* 5(3). 77–90.
- Kennedy, Claire; Miceli, Tiziana. 2010. “Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource.” *Language*

- Learning & Technology* 14(1). 28–44.
- Kilgarriff, Adam; Kosem, Iztok. 2012. “Corpus tools for lexicographers.” In Granger, S. & Paquot, M. (Eds.) *Electronic Lexicography*. 31–56. doi: <https://doi.org/10.1093/acprof:oso/9780199654864.003.0003>
- Lee, David; Swales, John. 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.
- Li Siu-Leung, Edward; Pemberton, Richard. 1994. “An investigation of students’ knowledge of academic and subtechnical vocabulary.” In Flowerdew, L. & Tong, A. (Eds.) *Entering text*. Hong Kong: Language Centre, Hong Kong University of Science and Technology. 183–196.
- Nation, Paul. 1990. *Teaching and learning vocabulary*. New York: Newbury House.
- Römer, Ute. 2008. “Corpora and language teaching.” In Lüdeling, A. and Kytö, M. (Eds.) *Corpus linguistics: An international handbook*. Berlin: Mouton de Gruyter. 112–130.
- Römer, Ute. 2011. “Corpus research applications in second language teaching.” *Annual review of applied linguistics* 31. 205–225.
- Tadić, Marko. 2003. *Jezične tehnologije i hrvatski jezik*. Zagreb: Exlibris.
- Todd, Richard Watson. 2001. “Induction from self-selected concordances and self-correction.” *System* 29(1). 91–102.
- Thurstun, Jennifer; Candlin, Christopher N. 1998. “Concordancing and the teaching of the vocabulary of academic English.” *English for Specific Purposes* 17(3). 267–280.
- Tribble, Christopher. 2004. “Managing relationships in professional writing.” In Hidalgo, E., Quereda, L. & Santana, J. (Eds.) *Corpora in the Foreign Language Classroom Selected papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6)*. Amsterdam and New York: Rodopi. 289–308.
- Vyatkina, Nina. 2016. “Data-driven learning of collocations: Learner performance, proficiency, and perceptions.” *Language Learning & Technology* 20(3). 159–179.
- Vyatkina, Nina. 2020. “Corpus-informed pedagogy in a language course: Design, implementation, and evaluation.” In Kruk, M. & Peterson, M. (Eds.) *New technological applications for foreign and second language learning and teaching*. Hershey, PA: IGI Global. 306–335.
- Wong, Lillian L. C. 2019. “Implementing disciplinary data-driven learning for Post-graduate thesis writing.” In Hyland, K. and Wong, L. L. C. (Eds.) *Specialised English: New Directions in ESP and EAP Research and Practice*. London and New York: Routledge. 195–213.
- Yoon, Hyunsook and Hirvela, Alan. 2004. “ESL student attitudes toward corpus use in L2 writing.” *Journal of Second Language Writing* 13. 257–283.