8 Software Development for Corpus Research in English Studies

Elizaveta Smirnova
HSE University, Perm

Svetlana Strinyuk
Admiral Makarov State University of Maritime and Inland Shipping, Saint Petersburg

Viacheslav Lanin
HSE University, Perm

In recent years, English has become the lingua franca of the spheres of higher education and science in Russia: more and more university courses are delivered in English, university students and academics take part in international conferences and workshops, and Russian scholars strive to publish their research findings in international peer-reviewed journals. Such a shift in focus has made the ability to write a high-quality academic text a necessary skill in the modern academic environment. However, as our experience as English for Academic Purposes (EAP) practitioners shows, Russian speakers writing in their second language (L2), having a good command of general English, often find it challenging to conform to the conventions of English academic discourse when writing their research papers or project proposals. Despite the existence of various types of software which can check the grammar and style of a text (e.g., Grammarly, Ginger, Language Tool), detect rhetorical moves in a text (Pendar & Cotos, 2008), and even provide feedback about errors (see, for example, Dreschler et al., 2019; Napolitano & Stent, 2009), to our knowledge, there are no programs focusing on the linguistic characteristics of an academic text. Besides, in the existing literature there appears to be no clear rubric for academic writing assessment. The application Paper Cat, developed by a team of teachers and students from HSE University, Perm, Russia, is aimed at facilitating students' and researchers' writing in English by

DOI: https://doi.org/10.37514/INT-B.2021.1428.2.08

> identifying the most significant features of academic discourse. We used the world-accumulated knowledge in EAP to develop a software that is able to assess an academic text against a set of criteria (i.e., academic discourse markers selected from academic style guides, handbooks and research articles on EAP). Evaluating the "quality of academic discourse" of the text in terms of style can be automated by using software to tag style markers in that text. At the heart of this approach is creating a repository of patterns which are needed to extract the markers mentioned above. The quality of L2 academic writing is assessed against a set of criteria based on an analysis of competent writing features.

English has become the lingua franca for the academic world (Drubin & Kellogg, 2012; Garfield, 1967; Meneghini & Packer, 2007). It dominates scientific literature, which means that a manuscript published in English immediately becomes more visible and significant (Drubin & Kellogg, 2012). Russian universities, being a part of the international academic community, strive to create an English-speaking environment to teach their students reading and writing skills in academic English by using English as a medium of instruction. Students find reading scientific literature and listening to lectures in English difficult, but exposure to the language in the educational environment does ultimately develop students' receptive skills. The battle which Russian learners of English typically have lost is with academic writing, which is demonstrated to some extent by the results of international exams: according to data from the official International English Language Testing System (IELTS) website, the mean overall and individual band scores achieved by 2019 Academic Training test takers of IELTS show that academic writing even in the simple form of a short essay and diagram description pose a serious difficulty for Russian students (see Table 8.1).

Table 8.1. Mean Band Score for the Most Common First Languages (Academic)

First language	Listening	Reading	Writing	Speaking	Overall
Russian	6.75	6.74	5.87	6.51	6.53
German	7.86	7.64	6.62	7.44	7.45
Italian	6.90	7.30	6.20	7.17	6.74
Tamil	6.87	6.43	5.98	6.53	6.52
Hindi	6.69	6.17	5.93	6.33	6.34

(International English Language Testing System, 2019)

The table shows that even though writing scores were lower than those for the other exam sections among all test-takers, Russian-speaking candidates have done worse in writing than speakers of other languages, even when they have a higher overall band (see, for example, the results of Hindi and Tamil-speaking candidates).

Therefore, as established in the first section of this book, increasing the level of students' English for academic purposes (EAP) writing skills has recently become a highly topical issue, since the ability to write a high-quality academic text is seen as a necessary skill in the modern academic environment. However, even those second language (L2) writers who have achieved a relatively high level of language proficiency have often found it challenging to conform to the conventions of academic discourse when writing their research papers in English (see Chapters 1, Chapter 3, Chapter 4, and Chapter 6). Writing in an appropriate academic style involves the use of particular lexical, grammatical, and syntactic structures associated with this type of discourse. Despite the existence of a large number of textbooks and study guides in academic English along with software which can check the grammar and/or style of a text (e.g., Grammarly, Ginger, Language Tool), teaching writing in a proper academic style remains a major challenge for EAP practitioners. To solve this problem, a research team from HSE University (thereafter HSE) in Perm, Russia has made an attempt to create a software that conducts a multidimensional analysis of academic English. We assume that the software can play an important role in analyzing academic discourse as well as in teaching English for academic purposes. The approach is based on data-driven learning (DDL; Johns, 1991, 2002), which involves giving learners access to language data to meet their learning needs. This approach uses large amounts of data (language corpora) in order to develop students' language skills and raise their stylistic consciousness. Using DDL in EAP classrooms has proven to be an effective way of developing learners' genre knowledge and discipline-specific writing skills (see, for example, Anthony, 2016; Cotos et al., 2017; Feak, 2016).

The main aim of this two-year project was to develop software capable of assessing an academic text against a set of criteria (i.e., academic discourse markers). The motivation behind the development of the software was to assist HSE students and lecturers with writing their papers in English.

Project Motivation and Development

In their final year, HSE students take a course in Academic Writing in English and write a research proposal as their final assessment. The research proposal is a draft of the students' diploma project written in English and edited according

to American Psychological Association (APA) and Institute of Electrical and Electronics Engineers (IEEE) style (depending on the major: social sciences or information technologies) and comprises roughly 2500 words. As teachers of this course who spend a great deal of time marking students' texts, we have concluded that the major difficulties students face in this work are not connected with content or grammar but with academic style in general (i.e., the use of lexical bundles and syntactic constructions expected in academic texts). During this short EAP course, they cannot master academic English at the necessary level. What is more, misleading instructions provided by handbooks in EAP which fail to clearly represent variations in conventions of academic English in different subject domains only add to the problem. For example, according to researchers, explicit evaluation through evaluative attributes has been more common in humanities and social sciences than in natural sciences, while modality as a way of expressing personal stance is more typical of natural sciences (Sotesbury, 2003). Clausal features occur more frequently in arts and humanities than life and physical sciences (Staples et al., 2016). However, these differences are not usually reflected in EAP textbooks.

Our software tool, developed using General Architecture for Text Engineering (GATE), is aimed at aiding students as they write. So far, learning programs have automatically detected rhetorical shifts (namely, establishing a territory; establishing a niche, occupying the niche) in academic texts (Pendar & Cotos, 2008); provided trigger questions and "gloss" (i.e., feedback content, which are supposed to help learners to reflect on and therefore improve their writing (Villalón et al., 2008); and identified and classified morphological and syntactic errors, suggesting ways of correcting them (Napolitano & Stent, 2009). Ours is different because 1) it is focused on academic discourse markers which are expected in advanced writing in a particular field; 2) it compares a user's text against a corpus of research articles in the same subject, which ensures a discipline-specific approach; 3) it uses statistics on the use of these markers, which contributes to the robustness of the assessment. Our tool identifies the most significant features of academic discourse within the subject domain based on corpus research and then uses that information to provide feedback to writers. It will also allow tutors to evaluate the quality of student papers against a number of standardized formal criteria.

The software also targets our colleagues who teach at HSE and are writing their own research papers in English for peer-reviewed journals. Writers could use this application to get real-time feedback during this challenging task. The application will be published as a publicly available service for comparing a user-provided text with text corpora. Since the program is based on GATE (Cunningham et al., 2011), which is free to use, the application is free as well.

GATE was chosen for several reasons. First, it provides a development environment (GATE Developer) with many basic processing resources (e.g., tokenizers, sentence splitters, morphological taggers) as well as an object library that can be used to write plugins specifically for the task (GATE Embedded). The main feature of GATE is a wide range of tools for text processing. The most useful tool for our project is the Java Annotation Patterns Engine (JAPE) transducer, which allows the user to describe regular expressions over GATE annotations. On the one hand, JAPE expressions can be used to find simple markers; on the other hand, we can write Java codes for complex markers ourselves.

The first version of the software tool was developed as a set of plugins for GATE Developer. Most of the plugins are used for finding style markers. At the same time, some plugins are aimed at statistical calculation and visualization. Based on acquired experience, we are now developing the second version of our tool as an internet research portal. Our portal will be able to perform a full circle of text processing from document and corpora management to building statistical reports. Due to its service-oriented architecture, the heterogeneous components of our solution can be seamlessly integrated together. Natural Language Processing (NLP) services are built on the GATE Embedded tool. Also, we have developed special tools such as a visual editor for JAPE expressions based on an ontological description. The portal can be used both for research and study aims.

We assume that evaluating the quality of the academic discourse of a text in terms of style can be automated by using software to tag style markers in that text. Creating a repository of patterns is at the heart of this approach, but it demands close attention. Therefore, at the first stage, it was necessary to make a list of patterns needed to extract the markers mentioned above. Evaluation of the statistical bounds of markers' occurrence requires using the methods and tools of corpus linguistics. In order to assess the quality of an academic text, the system compares it with a corpus of research papers published in leading peer-reviewed journals in different disciplines (i.e., a reference corpus).

So far, we have compiled 12 corpora—six of professional writing and six learner corpora in management, economics, history, political science, law, and computer science. The papers in the expert corpora were published between 2013 and 2020, and the sizes of the corpora and the journals the papers were retrieved from are presented in Table 8.2. Following Swales (1990), we believe that a paper published in a peer-reviewed journal can be seen as a model for L2 writers to follow, an academic text which "has a dynamic relationship" with various research-oriented genres, such as dissertations, monographs, presentations (Swales, 1990, p. 177). The research papers and research proposals written by HSE students have a similar macrostructure: they describe the

topic of the research, the knowledge gap, give a literature review on the topic, data, and methods, and present the results of the analysis (or anticipated results in some cases). It should be noted that a research proposal is the closest type of writing in English the student writers will do during their studies, because they go on to undertake and write up the research they proposed in their native language. Besides, the practice of comparing learner academic texts with professional writing is well established in EAP literature (see, for example, Aull et al., 2017; Lee & Chen, 2009; Smirnova, 2019). So, we believe that the corpora are comparable and can be used for our purposes.

Table 8.2. Sources of Texts and Sizes of the Expert Corpora

Discipline	Number of texts	Number of words	Journals
Economics	57	654,373	*Quarterly Journal of Economics* *Journal of Financial Economics* *International Journal of Production Economics*
Management	61	683,287	*Journal of Management* *Journal of Management Studies* *Academy of Management Journal*
Political Science	73	654,628	*American Political Science Review* *American Journal of Political Science* *Journal of Politics* *World Politics* *Comparative Political Studies* *Political Analysis*
Law	91	738,383	*European Law Journal* *European Law Journal* *Criminal Justice Studies* *Journal of Crime and Justice* *Contemporary Justice review*
History	65	621,723	*The American Historical Review* *The Journal of African History* *The Historical Journal* *the Journal of Modern History* *Contemporary European History*
Computer Science	86	705,271	*Artificial Intelligence Review* *European Journal of Information Systems* *Computer Science Education* *International Journal of Digital Earth*

Features that are used for the analysis are selected from academic style guides and other methodical literature (e.g., Hamp-Lyons & Heasly, 2010; Siepmann et al., 2011; Wallwork, 2016). They can be divided into three groups: lexical markers, grammar markers, and syntactic markers. The lexical markers include terminology, abstract semantic verbs, desemantisized verbs, intensifying adverbs, hedges, exemplification, and transition words. The grammar markers comprise the passive voice, present tenses, subject pronouns, and anaphoric expressions. The syntactic markers are pre- and postpositive attributes, it-clefts, pseudo clefts, non-finite clauses, adverbial clauses, th-wh constructions, and attitudinal clauses. It should be noted that the list is not full and is still being extended. A number of previous works (see, for example, Gray, 2015; Hyland, 2008; Staples et al., 2016) have demonstrated that there are significant disciplinary differences in the use of different lexical patterns and syntactic constructions in academic discourse. Therefore, the software we are developing is based on the discipline-specific approach.

User Experience and Application

Currently, our application offers three options: it is capable of annotating texts with the listed markers, providing statistics on their use, and assessing a user's text against a set of formalized criteria. Figure 8.1 shows logic connectors found by the software in an academic text.

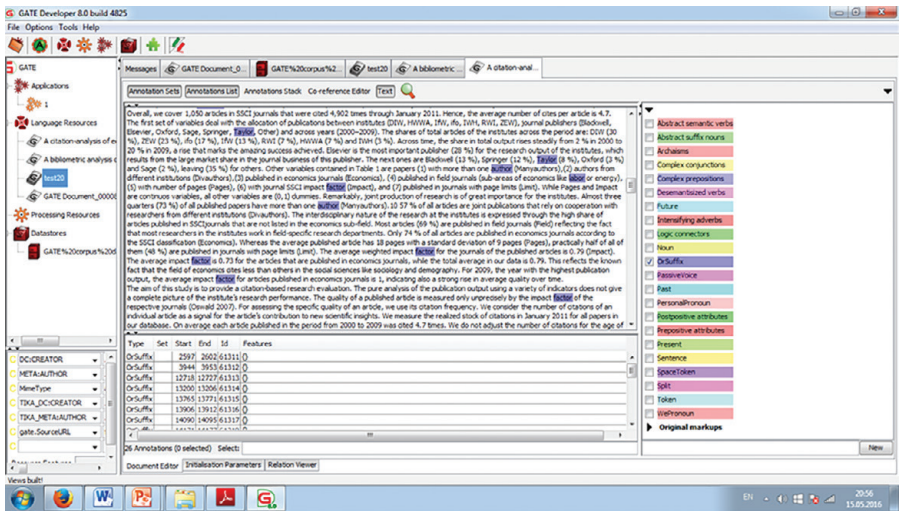

Figure 8.1. Logic connectors.

For the use of each marker, a student can get a maximum of ten points for the normal use of each marker of academic style. The norm is expressed quantitatively as the normalized frequency of a marker per thousand words in the reference corpus multiplied by the number of words in a student's work. A 10 percent deviation is possible for getting the maximum score. However, a larger deviation from the norm means a lower mark.

For example, the normalized frequency (occurrence per 1,000 words) of adverbial clauses in the reference corpus is seven. The work under consideration is 2,300 words long. Therefore, the usage norm for adverbial clauses for this work will be 7 times 2.3, which equals approximately 16. So, if there are 16+/- two adverbial clauses in the text, the student will get ten points for the use of this marker. If there are only ten adverbial clauses in the text, which is about 35 percent less than the norm, the student will get only 4 points. However, it should be mentioned that a tool like this should never actually be used for grading but possibly for self-regulation and formative feedback. This software is used in EAP classrooms at HSE in different ways.

Generating Study Materials

The compiled corpora of both expert and student writing can be used to generate study materials to assist in the classroom and in students' autonomous work. The use of concordances heightens the salience of linguistic units which the teacher or student wishes to focus on and thereby makes them more noticeable, which is a crucial factor in intake (Schmidt, 1990, 1994). Another benefit is that demonstrating concordance lines to learners can encourage them to process the material in a more profound way and to draw conclusions about the language units presented by themselves (Bernadini, 2004), thereby fostering learners' autonomy. Concordances showing the usage of some patterns aim not at providing students with answers but at giving them the tools for arriving at an independent solution to the problems they face when trying to express themselves in English (Johns, 2002). Moreover, presenting a lot of examples of a language feature in a concentrated way (i.e., in corpus lines), can save a lot of classroom time (Cobb, 1999; Hoey, 2000).

Expert writing corpora are extremely useful for creating various exercises as well as finding examples of the use of the identified markers of academic style. For instance, Table 8.3 demonstrates the most common uses of the hedging device *suggest* in the reference corpus of papers in management. Students can be asked to figure out the patterns of its use from some concordances on their own.

Table 8.3. Instances of *Suggest* in the Reference Corpus in Management

1.	*We*	***suggest***	*that, when a coercive pressure is introduced to adopt a new practice that is interpreted negatively by key institutional constituents.*
2.	*Our data*	***suggest***	*that decision makers take more time to comply with coercive pressures the more complexity they face.*
3.	*Trust between entrepreneurs and their investors has often been*	***suggested***	*to be key to their cooperation and the success of their partnership.*
4.	*A climate in which it is safe to speak up and take risks is*	***suggested***	*to complement the adaptation and implementation of innovation.*
5.	*The pattern of mediation that we uncovered*	***suggests***	*the possibility of other pathways such as affect.*
6.	*This*	***suggests***	*the potential for organizational interventions designed to bolster an individual's self-esteem level to potentially counteract ostracism's negative effects on self-esteem level.*

Based on the examples, learners are supposed to notice three patterns suggest is used in: somebody or something suggests that (1, 2); something is suggested to do something (3, 4); somebody suggests something (5, 6).

The learner corpora can also be used for generating error correction and text editing exercises. For examples, Table 8.4 shows examples of students' inaccurate use of anaphoric expressions, which can be employed for creating an error-correction exercise.

Table 8.4. Learners' Use of Anaphora

1.	Methods are effective only if it brings results in accordance with the goals and objectives.
2.	To obtain more specific information on each point I will select criteria and make a comparison of results according to it.
3.	The coach reflects the client's actions and helps to transform it into autonomous abilities (functions).
4.	The main idea in the third sub-group of corporate citizenship is that corporations can take its rightful place in society, next to other "citizens".
5.	The interview questionnaire will be developed on the basis of the reviewed literature, and it will help in gathering all the relevant data from this major informants.

Demonstrating Academic Discourse Markers in Use

The software can be used in an EAP classroom in order to demonstrate significant markers of academic style relevant to the discipline they are studying. For example, Figure 8.2 demonstrates the use of nouns and the passive voice in an academic text, which can serve as a colorful illustration for learners.

Moreover, the software can show typical discrepancies on the use of various markers in expert and learner corpora in a clear and catchy way (see Figure 8.3). This might allow teachers to prevent possible problems with the use of the markers in the future. The software can also be used to search for particular examples when studying certain markers of academic discourse.

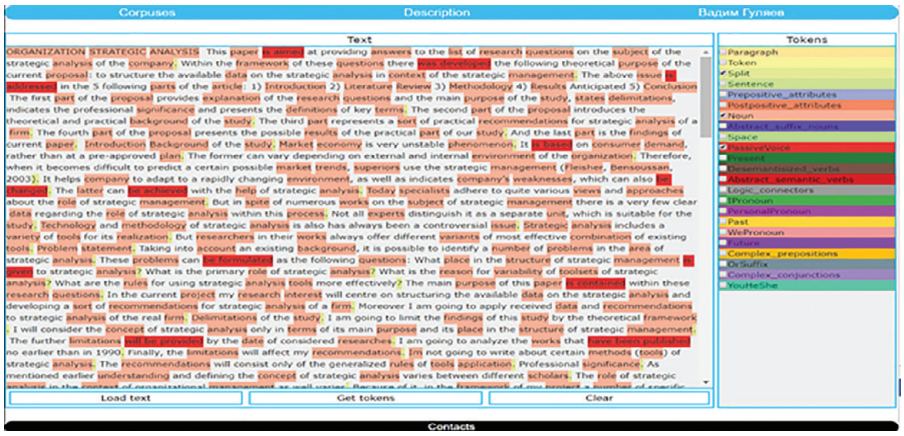

Figure 8.2. The use of nouns and the passive voice.

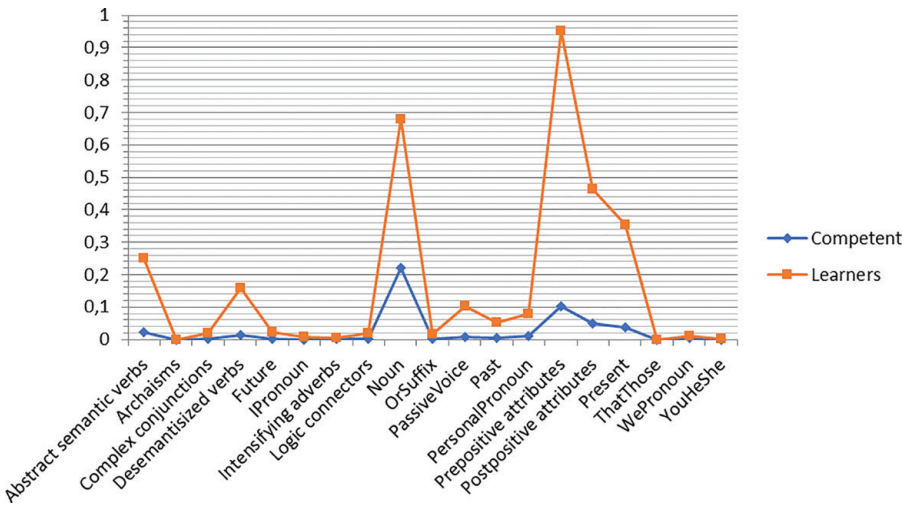

Figure 8.3. Comparison of learner and expert corpora by markers.

Motivating Learners' Autonomous Work

Finally, the software can be employed to motivate learners' autonomous work. The teacher can ask the students to find examples of some markers or identify some patterns of their use by themselves. Students are also able to upload their own academic text in order to get an automated, data-informed assessment of their work and subsequently try and improve it on their own.

While working on the project, we analyzed a lot of language data that allowed us to work out some practical recommendations that might be useful for EAP practitioners. According to our findings, not all syntactic features mentioned in EAP textbooks and study guides are frequently used by professional writers. Such rarely used syntactic constructions are, for example, it-clefts, pseudo-clefts, th-wh constructions, and adverbial clauses of purpose and manner. This might suggest that that under the conditions of limited classroom time, EAP teachers may exclude them from courses in academic writing or allocate the constructions to learners' self-study.

Conversely, our analysis informs us about which constructions ought to be prioritized depending on the learner's discipline. For instance, when teaching academic writing to learners in the hard sciences, it is important to allocate enough time for adverbial clauses of place, condition, and result because they are extensively used in comments for calculations, models, and formulas. For example:

1. Thus, the optimization formulation follows Eq. (4), where P and t are the decision variables.
2. If we apply the change of variables r we have that RIo xo; yo and, therefore, the Russell output measure of inefficiency is equivalent to an additive-type measure.
3. For convenience, we multiply the Amihud illiquidity measure by −1 so that the timing coefficient based on this measure has the same interpretation as that from the Pástor-Stambaugh liquidity measure.

On the other hand, courses in academic writing for learners who are studying soft disciplines should focus on adverbial clauses of time, contrast, and concession. It should also be mentioned that our research showed that learners do not use adverbial clauses of all types as frequently as professional writers do, sticking to simple sentences, for example:

4. Some of enterprises try to make and implement their own business processes models. Others try to use existing analysis and improvements models.
5. Logistics appeared in the Roman Empire. Its main task was the distribution of food.

This might suggest that this syntactic marker of academic style deserves close attention.

In contrast, complex sentences with which, who, and whose are used by students and professional writers with similar frequencies, but learners often use them incorrectly, for example:

6. External marketing of the employer brand is the second step that is needed to attract potential employees, which may become loyal employees in the future.

Another structure that was overused by novice writers, according to our research, was attitudinal clauses (in the analyzed learner corpora, they were used several times as often as in the expert ones). This finding suggests that students should be taught that these constructions can be used in their text once or twice only.

A particularly difficult structure for students turned out to be non-finite clauses; they are used by learners much more rarely than by professionals. This construction is quite frequent in the reference corpora and deserves special attention in EAP classrooms.

Anaphora also deserves the close attention of academic writing teachers because almost all types of anaphora have been underused by students. Thus, in students' works, there have been a lot of repetitions that could have been avoided with the help of anaphora, for example:

7. Although respondents are likely to be understood and to allow that such goods will be more expensive, but for ensuring environmental safety respondents agree to pay.
8. The first group of theories consists of utilitarian theories.

Our findings showed that the only type of anaphoric expressions overused by learners was demonstrative pronouns:

9. However, this paper supposes the use of customer development methodology for several reasons. Firstly, this method was adapted to IT-projects, for example, this technique involves the use of the approach of agile software development. Furthermore, this method is the least resource-consuming and it allows to test the hypothesis on a real market, using MVP. Finally, this technique was actually applied in practice in the majority of successful start-ups that participated in start-up accelerator of Russian internet Initiatives Development Fund.

The example demonstrates that students use it whenever they can, which might be due to an attempt to avoid errors related to the use of articles. This

implies that even though the use of articles is studied at a basic level in EFL courses, this topic should be revisited in EAP classrooms as well.

A large number of errors were connected with the use of plural nouns, even though this topic has also been studied at basic levels. Our experience shows that it requires repetition in academic writing courses in order to prevent possible errors, like the ones in the examples below.

10. To obtain more specific information on each point I will select criteria and make a comparison of results according to it.
11. This beliefs and expectations produce norms that powerfully shape the behavior of individuals and groups in the organization.

Reflexive pronouns require revision as well, according to our data:

12. Therefore, the manager not only itself has to adhere to ethical standards, but also has to provide their observance in the organization in general.

Hedges, which are seen as a lexical marker of academic style, have been generally underused by learners. Special attention should be paid to the use of the modal verbs may and might, which are rarely used in students' texts, along with the words seem and possible, which are abused by students.

As our corpus analysis suggests, students demonstrate incorrect usage of hedging devices, the most typical errors being related to the use of suggest with an infinitive:

13. The author suggests to approach the question from various perspectives of analyzing the market.

The use of the phrase "become possible," which is a word-by-word translation from Russian; and the use of several hedges together:

14. In future it seems possible to put theory into practice, and develop new technologies to establish a new business.

Conclusion

To conclude, writing style plays a pivotal role in presenting the results of research, and to be published, scholars have to meet the strict requirements of scientific journals. Researchers who are not native speakers of English struggle through manuals and guidelines for academic writing, but even so, materials are often rejected due to the low quality of the writing. Special courses in EAP or picking up language from academic papers are not always sufficient

remedies due to the natural limits of time and effort. Students taking English for Academic Purposes as a subject at university face an even more difficult situation due to strict deadlines and limited time in which to conduct a thorough literature review (which is practically the only type of academic reading they usually do). These limitations prevent them from processing and picking up the language in the most natural way—through extensive analytical reading. Digging in handbooks and manuals in EAP cannot sort out the problem since such studies are time consuming and usually require the assistance of a competent writer.

The application could solve this problem by assessing academic writing and providing those who would like to master academic writing with quality feedback. The application solves manifold tasks, namely, it may be used in corpus and contrastive research in order to analyze the L2 academic writing of both novice and competent writers; it can be used for teaching EAP in class and also for creating study materials. All these functions are fulfilled through the implementation of corpus research based on specific domain corpora which makes both teaching and researching more reliable.

References

Anthony, L. (2016). Introducing corpora and corpus tools into the technical writing classroom through data-driven learning (DDL). In J. Flowerdew & T. Costley (Eds.), *Discipline-specific writing: Theory into practice* (pp. 162-180). Routledge.

Aull, L., Bandarage, D., & Miller, M. R. (2017). Generality in student and expert epistemic stance: A corpus analysis of first-year, upper-level, and published academic writing. *Journal of English for Academic Purposes, 26*, 29-41. https://doi.org/10.1016/j.jeap.2017.01.005

Bernadini, S. (2004). Corpora in the classroom: an overview and some reflections on future developments. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 15-36). John Benjamins.

Cobb, T. (1999). Breadth and depth of lexical acquisition with hands-on concordancing. *Computer Assisted Language Learning, 12*(4), 345-360. https://doi.org/10.1076/call.12.4.345.5699

Cotos, E., Link, S., & Huffman, S. R. (2017). Effects of DDL technology on genre learning. *Language Learning & Technology*, *21*(3), 104-130. https://lib.dr.iastate.edu/engl_pubs/103/

Cunningham, H., Maynard, D., & Bontcheva, K. (2011). *Text processing with GATE (Version 6)*. University of Sheffield.

Dreschler G., Ontrust, M., & de Jong, N. (2019, July 1–4). *Digital individual support for writing skills across the disciplines* [Conference session]. The 2019 Conference for the European Association for the Teaching of Academic Writing. https://easychair.org/smart-program/EATAW2019/2019-07-03.html#talk:106283

Drubin, D. G., & Kellogg, D. R. (2012). English as the universal language of science: Opportunities and challenges. *Molecular Biology of the Cell, 23*(8), 1399. https://doi.org/10.1091/mbc.E12-02-0108

Feak, C. B. (2016). EAP support for post-graduate students. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (pp. 513-525). Routledge.

Garfield, E. (1967). English—An international language for science. *The Information Scientist, 76*, 19-20. http://www.garfield.library.upenn.edu/essays/V1p019y1962-73.pdf

Gray, B. (2015). On the complexity of academic writing: Disciplinary variation and structural complexity. In V. Cortes & E. Csomay (eds.) *Corpus-based research in applied linguistics: Studies in honor of Doug Biber* (pp. 49-78). John Benjamins.

Hamp-Lyons, L., & Heasly, B. (2010). *Study writing: A course in writing skills for academic purposes.* Cambridge University Press.

Hoey, M. (2000). A world beyond collocation: New perspectives on vocabulary teaching. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 224-245). Language Teaching Publications.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4-21. https://doi.org/10.1016/j.esp.2007.06.001

International English Language Testing System (2019). www.ielts.org

Java Annotation Patterns Engine (JAPE). (n.d.). Retrieved April 18, 2019 from https://gate.ac.uk/sale/tao/index.html#x1-2020008

Johns, T. F. (1991). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (Ed.), *Perspectives on pedagogical grammar* (pp. 293-313). Cambridge University Press. https://doi.org/10.1017/CBO9781139524605.014

Johns, T. (2002). Data-driven learning: The perpetual challenge. In B. Kettemann & G. Marko (Eds.), *Language and computers: Teaching and learning by doing corpus analysis. Proceedings of the fourth international conference on teaching and language corpora, Graz* (pp. 107-117). Rodopi.

Lee, D., & Chen, S. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing*, *18*, 149-165. https://doi.org/10.1016/j.jslw.2009.05.004

Meneghini, R., & Packer A. L. (2007). Is there science beyond English? *EMBO Reports*, *8*(2), 112-116. https://doi.org/10.1038/sj.embor.7400906

Napolitano, D. M., & Stent, A. (2009). TechWriter: An evolving system for writing assistance for advanced learners of English. *Calico Journal*, *26*(3), 611-625. https://www.jstor.org/stable/calicojournal.26.3.611

Pendar, N., & Cotos, E. (2008). Automatic identification of discourse moves in scientific article introductions. In Tetreault, J, Burstein, J., & De Felice, R. (Eds.), *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 62-70). Association for Computational Linguistics. https://www.aclweb.org/anthology/W08-0908.pdf

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129-158. https://doi.org/10.1093/applin/11.2.129

Schmidt, R. W. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. In J. H. Hulstijn & R. Schmidt (Eds.), *Consciousness and second language learning: Conceptual, methodological and practical issues in language learning and teaching, thematic issue of AILA review—Revue de l'AILA* (pp. 11-26). Free University Press

Siepmann, D., Gallagher, J. D., Hannay, M., & Mackenzie, J. L. (2011). *Writing in English: A guide for advanced learners*. UTB.

Smirnova, E. A. (2019). Referential coherence of academic texts: A corpus-based analysis of L2 research papers in management. *Journal of Language and Education*, *5*(4), 112-127. https://doi.org/10.17323/jle.2019.9688

Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, *33*(2), 149-183. https://doi.org/10.1177/0741088316631527

Swales, J. M. (1990). *Genre analysis: English in academic and research settings.* Cambridge University Press.

Villalón, J., Kearney, P., Calvo, R. A., & Reimann, P. (2008, July 1–5). Glosser: Enhanced feedback for student writing tasks. In *2008 Eighth Institute of Electrical and Electronics Engineers (IEEE) International Conference on Advanced Learning Technologies* (pp. 454-458). IEEE. https://doi.org/10.1109/icalt.2008.78

Wallwork, A. (2016). *English for writing research papers*. Springer.