Incorporating Data-Driven Learning Approach into English Language Teaching by Using Corpus-Based Tools

Shuping Huang(✉)

College of Chinese Language and Culture, Jinan University, Guangzhou 510610, China
Huangshuping@hwy.jnu.edu.cn

Abstract. Over the past few decades, information technology has been developing rapidly and playing an increasingly vital role in our society. There is no doubt that technology has already profoundly influenced education around the globe and it will continue to do so in the future, so how to embrace the change and integrate technology properly and effectively into language teaching and learning is an issue worthy of consideration. Computer corpora are collections of authentic language data that can be processed by computers, and today it is more convenient for teachers and students to gain access to corpora than before. In view of this situation, we attempt to adopt the data-driven learning (DDL) approach and utilize corpus-based tools in the teaching of English as a Foreign Language in the university setting in China. In this article, we first review the differences of the traditional PPP (presentation, practice and production) language teaching paradigm and the DDL approach to language teaching, and then demonstrate two examples of applying corpus-based tools to English teaching and learning. By learning to discover features of the target texts through observing and analyzing the language data and statistics offered by corpus-based tools, students' skills in autonomous learning and discovery learning are enhanced.

Keywords: data-driven learning (DDL) · corpus linguistics · corpus-based tool · English teaching and learning

1 Introduction

Information technology has been developing fast in today's world and has influenced people's life in almost every aspect. As technology can be a great tool to transform education, it is necessary for language teachers to consider how to take advantage of the advanced technology to facilitate language teaching and learning. With the development of computer sciences, corpora and corpus-based tools are evolving rapidly and has become more accessible than before. English, as one of the most widely used languages in the world, is learned as a foreign or second language in many countries, including China. Although in recent years some Chinese scholars [3, 4, 10, 11, 14] have done research on the pedagogical application of corpora and corpus-based tools in foreign language teaching, He [4] points out the application of corpora in the field of English teaching

P. Qi and Z. Chen (Eds.): ICBDIE 2023, AISR 178, pp. 713–720, 2024.
https://doi.org/10.2991/978-94-6463-238-5_93

in China "has relatively lagged behind." In order to facilitate the teaching practice of English as a Foreign Language (EFL) in the university setting in China, we incorporate corpus linguistic tools into the English classroom teaching based on data-driven learning (DDL).

2 The PPP Model vs. The DDL Approach

The PPP model is a traditional model for English language teaching. PPP is the acronym of presentation, practice and production. First, in the presentation stage, the teacher focuses on a particular form of language and presents it to the students; then in the practice stage, the teacher arranges practice activities for students to practice what has already been presented; finally, in the production stage, students are encouraged to produce the target form in a much more flexible way [13]. Although this dominant paradigm has some merits, some scholars have raised doubts about it. For example, Lewis [9] criticized the PPP model, saying it "is wholly unsatisfactory, failing as it does to reflect either the nature of language or the nature of learning."

The DDL approach is closely related to the computer corpus. As Leech [8] once said, "the student-centered paradigm of 'discovery learning' – or what Johns has called 'data-driven learning'– can scarcely be better exemplified than through the use of the computer corpus." According to Kennedy [7], a corpus "is a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description." A corpus provides the opportunity to measure tendencies and distributions across registers (and genres) of language [1]. Simpson-Vlach and Swales [12] hold that "corpus linguistics is essentially a technology." This technology can be applied to the teaching of English, because corpus linguistics "methodologically offers learners relevant and meaningful data—frequency distributions and actual patterns of English vocabulary and grammar as used in natural, authentic contexts" [2]. In other words, a corpus is a goldmine of naturally occurring linguistic data, and language teachers can explore and utilize it for language research and teaching.

The notion of data-driven learning (DDL) was put forward by Tim Johns, one of the pioneers advocating the use of corpora in teaching. Johns' [6] DDL approach encourages teachers to offer language data and corpus-based tools to students and help students learn how to use corpora and corpus-based tools to discover the linguistic rules on their own. Through this learning process, students can "become autonomous in their own acquisition of the target language" [2].

The DDL approach differs from the PPP model in several aspects (c.f. [2, 6, 14]). The PPP model is more teacher-centered while the DDL approach is more learner-centered. The PPP model is basically top-down and deductive while the DDL approach is mainly bottom-up and inductive. Teachers following the PPP model usually work as knowledge imparters while teachers adopting the DDL approach often work as guides and supporters. The students taught in the PPP model tend to become passive knowledge receivers while students taught with the DDL approach are likely to be active autonomous learners. In order to foster students' autonomous learning abilities and enhance their skills in discovery learning, we adopt the DDL approach.

3 Two Examples of Adopting DDL Approach in College English Teaching

In this section, we offer ideas of incorporating DDL approach to college English classroom teaching in China with two corpus-based tools. We take the teaching of the first text in the English textbook *An Integrated English Course 6* (2nd ed.) compiled by He Z. X. and Zhang C. B. [5] for example. Published by Shanghai Foreign Language Education Press, a top language education press in China, this textbook series has been well received in many Chinese colleges and universities.

The first text of *An Integrated English Course 6* (2nd ed.) is a narrative text entitled *A Class Act* written by Florence Cartlidge. It tells a story of the author's own experience at a school during the Second World War. Due to the poor economic conditions of her family at that time, she couldn't afford all the required school uniform. Every day at the assembly she had to be pulled out of line to stand on the stage as a bad example of what not to wear to school. She was deeply hurt by this daily humiliation. Then one day, her favorite teacher Miss McVee said to her in front of the whole class that she was the brightest and loveliest sight in this entire dreary school. These kind words uttered by Miss McVee were totally out of her expectation and thawed her frozen heart immediately. That day, Miss McVee taught her and the whole class a lesson of compassion and Miss McVee's kind words have inspired her ever since.

3.1 The Application of AntConc 4.2.0[1]

Before studying the text in details, the teacher can teach students how to use corpus-based tools to predict the general information of the text. Here we take AntConc for instance. AntConc is a free and user-friendly corpus tool created by Laurence Anthony of Waseda University in Tokyo. AntConc 4.2.0, which was released in 2022, is the latest version. People can use it to conduct corpus linguistics research and data-driven learning.

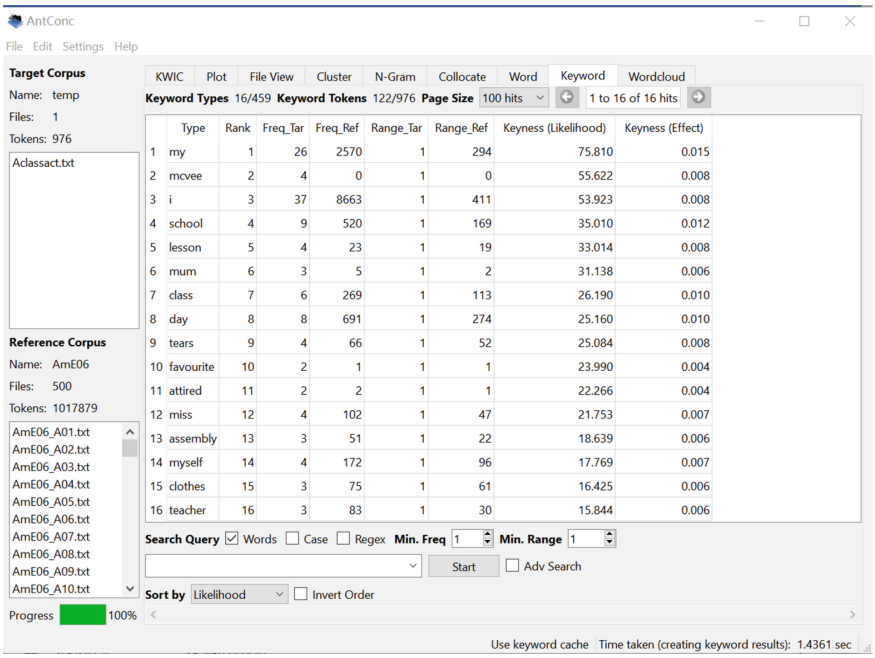
AntConc 4.2.0 has many functions. The teacher can use its Keyword tool function to help students predict the general information of the text. The Keyword list shows words that appear unusually frequently in the target corpus in comparison with the words in the reference corpus based on a statistical measure. These words can indicate the features of the target corpus.

Texts to be processed in AntConc 4.2.0 should be encoded in the UTF-8 encoding first. After encoding the target text *A Class Act*, the teacher clicks on the "Keyword" tool, and opens the encoded text as the target corpus for analysis, and then selects the AmE06 Corpus, a pre-built corpus database developed by Paul Baker available in the "Corpus Library" of AntConc 4.2.0, as the reference corpus, and after that clicks on "Start" at the bottom, and instantly a keyword list is shown in the main window (see Fig. 1).

There are sixteen keywords displayed in Fig. 1. They show the characteristics of the target file *A Class Act*. The sixteen keywords can be classified into four groups according to their parts of speech (see Table 1).

Then the teacher asks the students to make some predictions of the text by carefully observing Fig. 1 and Table 1. The purpose of the task is to inspire the students to explore

[1] AntConc 4.2.0 is available at https://laurenceanthony.net/software/antconc/.



	Type	Rank	Freq_Tar	Freq_Ref	Range_Tar	Range_Ref	Keyness (Likelihood)	Keyness (Effect)
1	my	1	26	2570	1	294	75.810	0.015
2	mcvee	2	4	0	1	0	55.622	0.008
3	i	3	37	8663	1	411	53.923	0.008
4	school	4	9	520	1	169	35.010	0.012
5	lesson	5	4	23	1	19	33.014	0.008
6	mum	6	3	5	1	2	31.138	0.006
7	class	7	6	269	1	113	26.190	0.010
8	day	8	8	691	1	274	25.160	0.010
9	tears	9	4	66	1	52	25.084	0.008
10	favourite	10	2	1	1	1	23.990	0.004
11	attired	11	2	2	1	1	22.266	0.004
12	miss	12	4	102	1	47	21.753	0.007
13	assembly	13	3	51	1	22	18.639	0.006
14	myself	14	4	172	1	96	17.769	0.007
15	clothes	15	3	75	1	61	16.425	0.006
16	teacher	16	3	83	1	30	15.844	0.006

Fig. 1. The keyword list of *A Class Act* by AntConc 4.2.0

Table 1. The POS of the keywords of *A Class Act*

POS	Keywords
Pronoun	my, I, myself
Noun	McVee, school, lesson, mum, class, day, tears, Miss, assembly, clothes, teacher
Adjective	favorite
Verb	attired

the data and discover some clues of the text. The teacher can also make the task less difficult by asking the students to consider the following questions: who may be the main characters of the text? where may the story happen? what event is probably being talked about in the text?

The teacher can guide students to make the following predictions: this is probably a story of the author's own experience, for the first person pronouns "my", "I", and "myself" are all in the keyword list; McVee can be a very important character in the story; the setting of the story is probably at a school, as most of the nouns in the keyword list are related to school life; there may be something unpleasant happening to someone, because "tears" appears in the keyword list; the text may be related to clothes, because "clothes" and "attired" appear in the keyword list.

After that, the teacher helps the students to generalize some rules of predicting the general information of a target text: in most cases in the keyword list, the personal pronouns and proper nouns especially people's names indicate the main characters in the target text, and the common nouns and verbs probably reveal the main events and the adjectives may show the attitudes or feelings of people.

In the end, the teacher can guide students to summarize the steps of making predictions about a text by using the corpus tool AntConc 4.2.0: first, encode the target text in the UTF-8 encoding and load it into AntConc as the target file; second, use the "Keyword" function of AntConc to get a keyword list; third, classify the keywords according to their parts of speech; finally, observe and predict the general information of the text.

Through the whole teaching and learning process, the teacher works as a guide and supporter, showing students how to use corpus-based tools for discovery learning and giving them support when they encounter difficulties.

3.2 The Application of AntWordProfiler 2.0.1[2]

AntWordProfiler is another freeware designed by Lawrence Anthony for profiling the vocabulary level and complexity of texts. AntWordProfiler 2.0.1, which was released in 2022, is the latest version available now. It has three built-in level lists created by Paul Nation and cleaned by Laurence Anthony, namely, 1_gsl_1st_1000.txt, 2_gsl_2nd_1000.txt, 3_awl_570.txt. The first list and the second list include 2, 000 most frequent word families from the General Service List (West, 1953) and the third list includes 570 word families from the Academic Word List (Coxhead, 2000). Users can also change the lists or add an "ignore list". AntWordProfiler can provide the statistics of the word coverage of the target files in the level lists and it can be used as an alternative to the Range software created by Paul Nation.

Teachers and students can use AntWordProfiler 2.0.1 to discover which level list each word in the target text belongs to and which words in the target text are not in the default list. Generally speaking, for language learners, the words covered in the 1_gsl_1st_1000.txt and 2_gsl_2nd_1000.txt are easier ones while words covered in 3_awl_570.txt or not in any of the three lists are more difficult ones.

In order to identify which words are the difficult ones in the text *A Class Act*, the teacher loads it as the target file, and double clicks on the name of the file "A Class Act", and the profiling of the text appears on the screen at once (see Fig. 2).

In Fig. 2, there are some words in the target text highlighted in orange. They are the words that are not covered in the default lists. Usually, they are also the words that students find unfamiliar or difficult. Teachers should spend more time teaching students these words. As the proper nouns are not added in the ignore list, "Florence Cartlidge", the author and "Manchester", the city where the story happened, are also highlighted in the text. If users don't want those words to be highlighted, they can just add them to the ignore list.

In the lower left corner in Fig. 2, the statistics of the word coverage are presented. 89.70% of the words in *A Class Act* are covered in the three lists while 10.30% are not in the reference lists or ignore list. The word coverage of each level list is also displayed.

[2] AntWordProfiler 2.0.1 is available at https://laurenceanthony.net/software/antwordprofiler/.

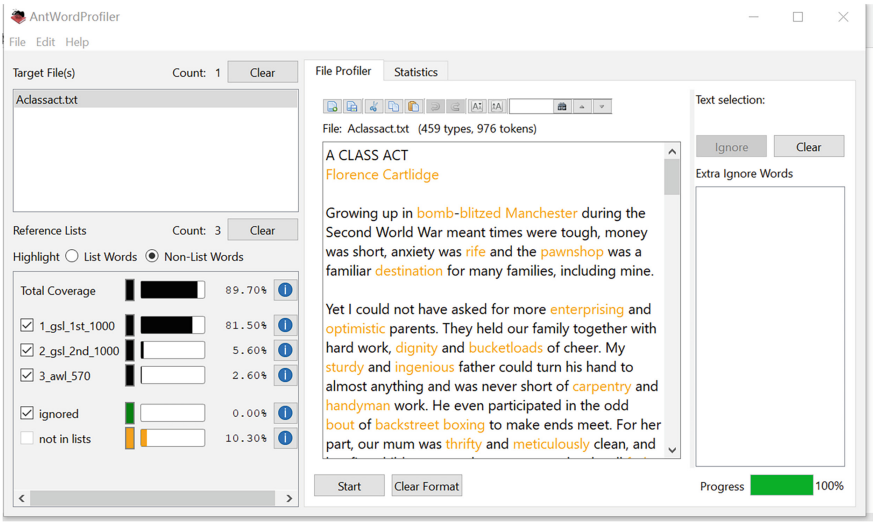
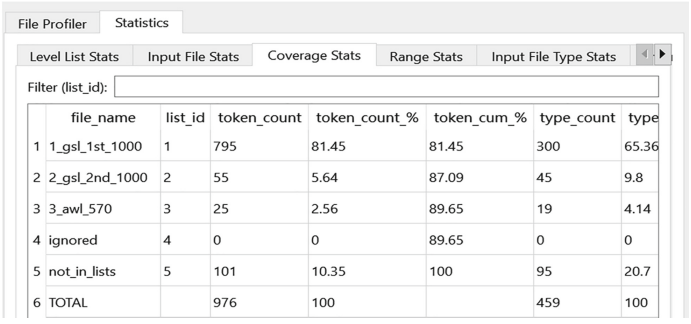


Fig. 2. The profiling of *A Class Act* by AntWordProfiler 2.0.1

Click on the blue information button (i) next to the percentage can show the specific words that appear in each level.

To see more detailed statistics, click on "Statistics" next to "File Profiler" and then click on the "Update Statistics" button at the bottom. In a second, statistics appear on the screen (see Fig. 3). There are different sub-tabs of statistics (e.g., Level List Stats, Input File Stats, Coverage Stats, Range Stats, etc.). Users can click on these sub-tabs to get the corresponding tables of statistics.

According to statistics in Fig. 3, 81.45% of the words in *A Class Act* are covered in the first GSL list; 5.64% of the words are in the second GSL list; 2.56% of the words are in the AWL list and the rest 10.35% of the words are not in any of the above list. Based on the above statistics, the teacher can infer that students may understand the general idea of the text, but they need to spend more time learning the 10.35% Non-list Words to understand more details of the text.



	file_name	list_id	token_count	token_count_%	token_cum_%	type_count	type
1	1_gsl_1st_1000	1	795	81.45	81.45	300	65.36
2	2_gsl_2nd_1000	2	55	5.64	87.09	45	9.8
3	3_awl_570	3	25	2.56	89.65	19	4.14
4	ignored	4	0	0	89.65	0	0
5	not_in_lists	5	101	10.35	100	95	20.7
6	TOTAL		976	100		459	100

Fig. 3. Part of the coverage statistics table of *A Class Act* by AntWordProfiler 2.0.1

To sum up, with AntWordProfiler 2.0.1 teachers and students can have a better idea of the vocabulary level and complexity of the target texts, which means it can be applied to English teaching and learning at least in two ways. First, it allows teachers and students to discover and focus on the difficult words of the texts, which usually belong to the third list (i.e., 33_awl_570.txt) and the fifth list (i.e., not_in_lists). Second, it can help teachers to select after-class supplementary English reading materials that are appropriate for the students' English proficiency, or students can use it to choose appropriate reading materials for themselves.

4 Conclusion

Today, technology has become an essential part of our life and has been constantly changing and even shaping the world to some extent. It is advisable for teachers and students to open arms to welcome the technological challenge. Computer corpora or corpus linguistic tools are useful in the field of language teaching and learning, as the authentic linguistic data of the corpora and statistics calculated by corpus-based tools allow for more objective language studies. In this article, we have presented two examples of using corpus-based tools to enhance college English teaching and learning based on the data-driven learning (DDL) approach. To be specific, in Sect. 3, we have demonstrated how to teach students to predict the general information of a text with AntConc 4.2.0 and judge the vocabulary level and complexity of a text with AntWordProfiler 2.0.1. Using the corpus-based tools can help students learn to observe and discover the features of the target texts from the linguistic data and statistics on their own, thus enhancing their autonomy and discovery learning abilities. Here we only present the application of two corpus-based tools for English teaching and learning. In the future, we will conduct more empirical studies on the DDL approach to English teaching and learning, and explore more pedagogical applications of easy-to-use corpus-based tools.

References

1. Friginal, E., Waugh, O., Titak, A.: Linguistic variation in Facebook and Twitter posts. In: Friginal, E. (ed.), Studies in corpus-based sociolinguistics, pp. 342-362. Routledge, New York (2017)
2. Friginal, E.: Corpus linguistics for English teachers: new tools, online resources, and classroom activities. Routledge, New York (2018)
3. He, A. P.: Corpus linguistics and English teaching. Foreign Language Teaching and Research Press, Beijing (2004)
4. He, A. P.: An introduction to corpus-assisted English teaching (rev. ed.). Foreign Language Teaching and Research Press, Beijing (2017)
5. He, Z. X., Zhang C. B.: An Integrated English Course 6 (2nd ed.). Shanghai Foreign Language Education Press, Shanghai (2013)
6. Johns, T.: From printout to handout: grammar and vocabulary teaching in the context of data-driven learning. In: Odlin, T. (ed.), Perspectives on pedagogical grammar, pp. 293–313. Cambridge University Press, Cambridge (1994)
7. Kennedy, G.: An introduction to corpus linguistics. Foreign Language Teaching and Research Press, Beijing (2000)

8. Leech, G.: Teaching and language corpora: a convergence. In: Wichmann, A., Fligelstone, S., McEnery, T., Knowles, G. (eds.), Teaching and language corpora, pp.1-24. Longman, New York (1997)
9. Lewis, M.: Implications of a lexical view of language. In: Willis, J., Willis, D. (eds.), Challenge and change in language teaching, pp.10-16. Shanghai Foreign Language Education Press, Shanghai (2002)
10. Liu, P.: Corpus-based data driven learning of English for academic purposes. Wuhan University Press, Wuhan (2021)
11. Liu, X. Q.: A multi-dimensional exploration of corpus-aided EFL autonomous learning. South China University of Technology Press, Guangzhou (2013)
12. Simpson-Vlach, R., Swales, J.: Corpus linguistics in North America: selections from the 1999 symposium. University of Michigan Press, Ann Arbor, MI. (2001)
13. Skehan, P.: Second language acquisition research and task-based instruction. In: Willis, J., Willis, D. (eds.), Challenge and change in language teaching, pp.17-30. Shanghai Foreign Language Education Press, Shanghai (2002)
14. Zhen, F. C.: Corpus-based data-driven foreign language learning and teaching: ideas, methods and technology. Foreign Language World. (4),19–27+40 (2005)