Data-Driven Learning in Enhancing Learners' Language Idiomaticity

https://doi.org/10.3991/ijet.v15i23.19023

Haiyan Men
Shanghai Sanda University, Shanghai, China
haiyanmen@gmail.com

Abstract—As data-driven learning has been advancing across new frontiers in recent years, there is still a paucity of studies on data-driven vocabulary learning model that brings about effectiveness in teaching and learning practices. Idiomaticity, which serves as an important indicator for language proficiency, needs abundant contextualized language input for the acquisition of target words. In this regard, the present study explores whether computer-assisted language learning is effective in vocabulary acquisition, and in the differentiation of synonymous words on the part of learners. Pre-/posttests and questionnaires were administered among an experimental group (N=26) and a control group (N=26). Results showed that the experimental group, who was instructed under the data-driven learning model, got a higher mean score than the control group, who received traditional dictionary-consulting instruction. The former also finished the posttest within a much shorter period of time. A significant relationship was found between the pretest scores and posttest scores among the experimental group whereas the scores in the control group did not reach statistical significance. Therefore, there was a significant improvement in learners' performance on collocation production under the data-driven learning model, whilst dictionaries did not prove to have such a contributing effect. This study provides some suggestions for how to enhance learners' idiomaticity by improving collocation performance under the data-driven model.

Keywords—Data-driven vocabulary learning, idiomaticity, collocations, COCA

1 Introduction

Vocabulary is the building block of a language. As a celebrated quotation goes, without grammar very little can be conveyed; without vocabulary nothing can be conveyed [1]. Vocabulary teaching in an English as a second or foreign language setting has long been occupying a central position in second language classrooms. With the development of computer technologies and educational sciences, vocabulary pedagogical approach has undergone major shifts in recent years. The last few years have seen a rapid development in computerized linguistic resources for the acquisition of L2 vocabulary. Computer-based language learning, or data-driven learning (DDL), has revolutionized traditional ways of learning. Traditional instruction has gradually

been proved to be of limited effects compared with technological pedagogies [2]. Under the DDL model, learners are not only provided with ample contextualized authentic language use data, which is a guarantee for quality language input, but also develop independent learning habits through generalizing language rules by themselves. Data-driven learning enables learners to discover to learn and to acquire language at any time, as the data provided to them is like "portable teachers". However, there is still a paucity of empirical evidence on data-driven vocabulary learning. Most studies on vocabulary acquisition have found learners' difficulty in learning L2 vocabulary, but only suggest contextualized DDL in the pedagogical implication parts. In addition, both teachers and students still fail to embrace DDL as they are not well aware of its benefits and thus do not benefit from its advantages [3]. In this regard, this paper tries to fill this gap by developing a data-driven vocabulary learning model, whose effects on vocabulary acquisition are further empirically explored. One most important component of word knowledge is which words the target word can go with, or knowledge of word combinations. A good command of word combinations can efficiently improve fluency and native-like production. So, this study aims to see if data-driven vocabulary learning enhances language learners' idiomaticity.

Idiomaticity serves as an important indicator for language proficiency. Multi-word units like collocations, fixed expressions, prefabricated chunks and idioms are common manifestations of idiomaticity. As in the process of first-language production, complete freedom choice of single words is rare and instead meanings are created through word combinations [4]. Both the written and spoken forms of a language are made up of large proportions of (semi-)fixed word combinations. For EFL learners, phraseological knowledge not only facilitates efficient comprehension, but also promotes native-like production.

Collocation is one of the most frequent and important forms of idiomatic uses of language. Given its nature of arbitrary restriction of word combinations, collocation poses particular difficulty for second language learners. L2 learners' collocational deficiency was reported as early as in the 1930s. EFL learners find word combinations that cannot be clearly explained by grammatical rules and are habitually used by native speakers most difficult to learn. For example, strong coffee is an acceptable English collocation, but learners may associate powerful with coffee, which is not an idiomatic English expression. Another example is major catastrophe, where learners may produce any pairings of words with similar meanings: big, large, important, and considerable, with disaster, calamity, mishap, tragedy, and the like [5]. Grammar rules in this regard are too general to delimit the acceptability of accurate word combinations. Collocation learning is thus identified to be a problematic domain for EFL learners. For example, according to [6], 108 Spanish learners of English only got a mean score of 56.6% when tested on their productive knowledge of 50 collocations. Collocation learning is reported to lag behind other aspects of second language acquisition [7]. Even with the rise in learners' proficiency level, the number of erroneous collocations rises accordingly [8]. One factor associated with the difficulty in learning L2 collocations, as found in empirical studies, is the misuse of semantically-related words, or synonymous words [9]. To put it more specific, the increasing lexemes in a synonym set learners have acquired are the main factor for the collocation lag. Psy-

cholinguistic studies have also found that learners have difficulties distinguishing word meanings as they have same-translation pairs stored nearby in the mental lexicon [10]. Thus, the acquisition and differentiation of synonymous words are of great importance for the attainment of native-like proficiency.

When it comes to the distinction of synonymous word pairs, the received wisdom years ago was to turn to dictionaries and thesauri. Dictionaries normally provide lexical information like word senses, typical collocates and example sentences, etc. For the general comprehension of word semantics, the information given by dictionaries is helpful. However, without large amount of authentic language use data, they fail to provide nuances of meanings of synonyms. In addition, the deduction learning method with which learners first memorize the meanings of a given word and then learn to use it in different tasks produces limited effect for them to retain knowledge. That's one of the reasons why traditional vocabulary instruction is beginning to receive criticism. In empirical studies on the role of dictionaries, it was found that over 36% of the learners investigated couldn't find the required information for synonyms through searching for word definitions in dictionaries [11]. In this regard, there remains a great need for providing learners with large amount of contextualized language use data, by which learners can observe and at the same time generalize word senses and uses. With the emergence of a considerable variety of corpora−collections of naturally-occurring language texts, data-driven vocabulary learning has begun to be widely advocated.

Corpus-based lexicology and vocabulary learning have long been endorsed [12, 13]. Yet it seems that both teachers and second language learners are not well aware of such a data-driven learning model. Students were found to know nothing about online corpora and concordances [14]. With the help of abundant authentic learning materials and user-friendly online tools, data-driven vocabulary learning can optimize learning effects, reinforce independent learning behavior and enhance students' problem-solving abilities. Therefore, from an applied point of view, it is useful to develop a model of data-driven vocabulary learning, in order to help learners fully acquire the semantics of the target words, to distinguish synonymous words in authentic language use on their own and to improve language idiomaticity.

2 Data-Driven Vocabulary Learning Model Design

The emergence and thriving development in corpus technologies, in combination with rapid developments in internet technology, provide learners with easy and quick access to abundant real and authentic language data. This convenience enables learners to generalize language patterns through observing language use. As an important approach for vocabulary acquisition, the inductive learning method also helps learners to discover to learn, and thus learners can develop independent learning habit.

The well-known dictum − "You shall know a word by the company it keeps" [15] has formulated practical guidelines for a data-driven vocabulary learning model. Unlike traditional vocabulary learning where students learn L2 words in isolation, data-driven learning enables learners to examine the "the company the word keeps", i.e., collocates. The data-driven learning tool applied in this study is the large-scale online

corpus-COCA (Corpus of Contemporary American English). Developed by Prof. Mark Davies of Brigham Young University, the corpus contains more than one billion words of texts covering a wide range of genres: spoken, fiction, popular magazines, newspapers, academic texts, TV and movie subtitles, blogs, and other web pages. The size of the corpus is still growing every day. The large number of authentic language use provides learners with an effective way to examine the English vocabulary. Meanwhile, a user-friendly interface is available for either learners or researchers to search for words, phrases, and strings, and to look for information regarding synonyms, concordances and collocates for a search word. COCA has been highly recommended by researchers as a good platform for English teachers and learners. Up to now, it has yielded many fruitful outcomes in the field of second language acquisition research.

Next, we take the data-driven learning of synonym pair - compliment and praise as an example. From the perspective of semantics, both words include meanings of approval and admiration. In terms of transitivity, both verbs are transitive and thus can be directly followed by nouns and noun phrases as objects. From the point of contrastive semantics, both have the same translation equivalents in Chinese. Therefore, without further lexical information, it is not an easy task for learners to distinguish their uses. COCA, in this regard, provides an effective way for meaning differentiation and vocabulary acquisition. Under the data-driven vocabulary learning model, the first step is to search for the verb compliment_v* in COCA, and then click "see detailed information for word". Then the corpus presents a wealth of information about the search term, as is shown in Fig. 1, 2 and 3.

Fig. 1. Search results for *compliment*

Fig. 2. Clusters of *compliment*

Fig. 3. Concordance display of *compliment*

As the three figures show, the web generates a series of categories of lexical information regarding the search term *compliment*. It presents the frequency distribution of genres where the word is usually used, its meanings, related topics, collocates (automatically grouped by parts of speech) and synonyms. In addition, the most frequent two, three and four strings of clusters of *compliment* and its concordances are also displayed. These entries enable learners to have a general understanding of the semantics and use of *compliment*. For example, the semantic information entry clearly explains the meaning of *compliment*: say something to someone that expresses praise; express respect or esteem for. *Flatter* is on the top list of its synonymous verbs, which suggests that *compliment* can be derogatory in meaning. Once learners acquire its semantic meaning, the next important step is to examine the use of the word and summarize its usage patterns. The cluster and concordance display (keyword in context display) in Fig. 2 and 3 provide ample resources for learners to generalize the use of *compliment*. The verb is usually followed by a pronoun and a preposition and the typical colligational pattern is *compliment somebody on something*.

In a similar vein, learners can formulate the usage patterns of *praise* through COCA. The collocational and colligation patterns for *praise* is *praise something/ praise somebody for* and its synonymous verb is *admire*, which is commendatory. Therefore, with similar meanings, the synonym pair *compliment* and *praise* has distinct differences in the shades of meanings, and thus has different uses (see Table 1 for the usage patterns of the two words).

Table 1. Usage patterns of compliment and praise

Words	Collocational / colligational patterns
Compliment	Compliment somebody on
	Synonym: flatter (Derogatory)
Praise	Praise something / praise somebody for
	Synonym: Admire (commendatory)

At the first stage, teachers can instruct students how to use COCA, and how to interpret lexical information and summarize the uses of the search terms. At later stage, prior to the development of the capacity for self-directed learning, students are encouraged to perform the above tasks under the guidance of teachers. At the last stage, after they are familiar with the procedures, they can conduct autonomous learning of the target words through using collocation information to observe word senses. It is worth noting that a higher English proficiency level is required at the independent learning stage, as learners need sufficient vocabulary knowledge in making generalizations.

3 Methodology

The study involved fifty-two first-year Chinese English majors studying at Shanghai Sanda University. The subjects came from two classes, one of which was the experimental group, and the other one was the control group. Their numbers were evenly distributed. The experimental class was instructed in the data-driven learning model and the control class was asked to distinguish words with the traditional aid of dictionaries (they were free to choose any dictionaries). Both groups had the same proficiency level, as their mean scores in the mid-term and final term examinations for the reading courses in each semester were similar. Besides, they got similar mean scores for the quizzes given to them every two weeks. The experiment was carried out at the end of the second semester of their first year, as by this time they had already received intensive training of the English language skills for one year and thus developed independent learning ability to some extent. The forms of the experiment made in this study included questionnaires, pretests, and posttests. Both quantitative and qualitative analyses were performed on the data collected.

3.1 Questionnaires

Subjects were asked to take part in two short questionnaires developed for investigating their attitudes towards and perceptions of collocation use and the online corpus learning tool. The questionnaire administered for both groups before the pretests included the following questions:

1. Do you find collocations difficult to use?
2. When distinguishing synonymous words, what reference tools do you usually turn to?

The questionnaires administered after the learning of the tested vocabulary through COCA and the use of dictionaries were to obtain feedbacks from students on the data-driven learning method and traditional vocabulary learning. They asked students if they found the tools (COCA and dictionaries) presented to them useful for differentiating synonymous words and producing correct collocations, and asked them to put the helpfulness on a scale of "very helpful", "helpful", "not very helpful" and "not helpful". Another question asked about their preferences when learning vocabulary, i.e., whether to use the online tools or dictionaries to carry out future independent learning.

3.2 Pre /posttests

Students were pretested on their knowledge of synonymous words and their collocations. After the data-driven learning model and the traditional learning method were adopted, posttests were administered to observe if their mastery of the tested vocabulary improved. Both the pretest and posttest consisted of the same set of twenty multiple choice items. The synonym pairs were selected from students' reading course textbook: *An Integrated English Course*. In order to strengthen students' ability in differentiating words with similar meaning, the textbook designed blank-filling exercises following each unit. Yet no effective ways were suggested by the textbook as to how to help students make distinctions between the synonym pairs. As verb + noun collocations are the most frequently used and most important type of collocations, 5 verb pairs were chosen: *compliment* vs. *praise*, *keep* vs. *maintain*, *demand* vs. *ask*, *shape* vs. *form*, *fulfill* vs. *realize* (see Appendix for the test, 20 items in total). The two tests were administered both online and students were asked to submit them in seven minutes. When finishing the tasks, they were not allowed to consult any dictionaries nor use the internet for searching the verbs. The seven-minute period also made it nearly impossible for learners to turn to dictionaries for help. The original blank-filling items were changed to multiple choices so that learners' main attention was directed to the differentiation of words meanings, rather than other grammatical rules. Another advantage of this test format is to utilize the auto-checking function of the online teaching platform − the Wisdom Tree. Students' scores were automatically calculated immediately after the tests. The platform could also calculate the accuracy rates for each item. SPSS software was used to analyze the data.

3.3 Procedures

The following figure vividly presents the steps involved in this study.

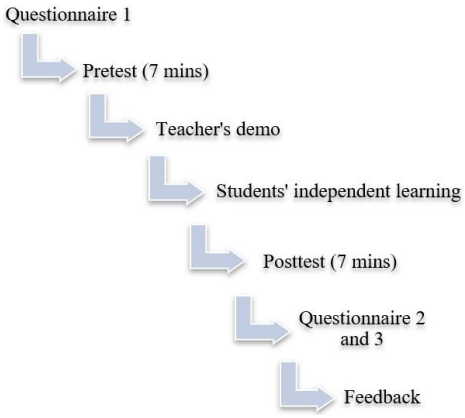

Fig. 4. A flow chart for the experiment procedures

As is presented in the Fig. 4, data were collected after the first six steps were taken. The last step was designed to give the students accurate and timely feedback on the tests. For the control group, keys to the blank-filling exercises and explanations were given; problems as to why they failed in choosing the correct verb from synonym pairs even after they had consulted dictionaries were addressed; for the experimental group, an additional feedback concerning the collocation/colligation patterns of the remaining four verbs (the compliment/praise pair was illustrated in the demo session) was given. This step aimed to let students check by themselves if their generalization was complete and accurate. Teachers could additionally check if subjects' independent learning reached a desired stage. As was shown in Table 1 in Section 2, the usage profiles of the four verb pairs were presented to participants (see the following four tables below).

Table 2. Usage patterns of keep and maintain

Words	Collocational/colligational patterns
Keep	keep it; keep doing something; keep an eye on; keep in mind
	synonym: continue
Maintain	maintain + level, control, relationship, balance, status, etc.
	synonym: uphold

Table 3. Usage patterns of demand and ask

Words	Collocational / colligational patterns
Demand	Demand + supply, attention, answer, etc.; Demand to know…; Demand that…
	Synonym: require
Ask	Ask + question, permission; Ask somebody about; Ask if…
	Synonym: request

Table 4. Usage patterns of shape and form

Words	Collocational / colligational patterns
Shape	Shape + future, policy, role, experience, culture, etc.
	Synonym: influence
Form	Form + group, basis, alliance, opinion, relationship, etc.
	Synonym: develop

Table 5. Usage patterns of fulfill and realize

Words	Collocational / colligational patterns
Fulfill	Fulfill + promise, need, obligation, dream, requirement, role, duty, etc.
	Synonym: carry out
Realize	Realize + potential, dream, ambition, etc. Realize that…
	Synonym: reach

4 Research Findings

4.1 The before and after tests

The test results of the experimental and control groups are shown in Table 6. Table 6 shows the summarization of the group performances, with respect to two ways of vocabulary learning in differentiating synonyms and producing the correct collocations.

Table 6. Test results display

Project		Pretest	Posttest
Experimental class	Mean	13.54	17.42
	Standard deviation	2.929	2.176
Control class	Mean	13.15	14.15
	Standard deviation	2.541	3.738

From the descriptive statistics, the following observations are made:

1. In terms of pretest scores, the two groups have similar scores (13.54 and 13.15, out of 20), which means that there is not a major difference in their English proficiency level. This increases the validity for the comparison of their performance in the posttests. The mean score also indicates that the accuracy rates of subjects' collocation uses is about 65%, and they do have considerable difficulty in producing the correct collocations. The score is similar with [16]'s study, where English major freshmen were reported to have a test score of 6.15 (out of 10) in collocation tests.
2. Subjects score higher in the posttests, meaning there is an improvement in the acquisition of the tested vocabulary, either with the data-driven learning method or dictionary-looking up method. However, the experimental class got a higher score (17.42) than the control class (14.15). This score difference shows that the data-driven learning vocabulary model is more effective than the traditional method. Conducting synonym distinction with the help of dictionaries turns out to produce very little effect.
3. As the standard deviation figure shows, subjects in the experimental group have a more reliable performance after the implementation of the data-driven learning model (from 2.929 to 2.176). Between-group comparison of the standard deviation indicates that subjects in the control class do not have an equally reliable performance in the tests, as the score (3.738) is the highest among all tests. Therefore, we can infer that subjects' consultation of dictionaries may produce confusing or even counteractive effects for some subjects on synonym distinction.

In order to see if the pretest and posttest scores in each group have statistical difference, a T test was performed on the above results. Pairwise comparisons were respectively made between the scores of the two groups. For the experimental group, the P value is .000 ($P<0.01$) and 0.172 ($P>0.05$) for the control group. A significant relationship is found between the pretest scores and posttest scores among the experimental group whereas the scores in the control group do not reach statistical significance. So, there is a significant improvement in learners' performance on collocation production under the data-driven learning model, whilst dictionaries do not prove to have such a contributing effect.

In addition, the time duration in taking the test questions was also investigated in this study, as it is of equal significance to see whether the data-driven learning model helps build students' confidence in doing the tests and whether it raises their efficiency. Subjects were asked to finish the 20 items within 7 minutes in both the pretests and posttests. The online teaching platform recorded the time when students began to take the tests and the time of their submission. For the experimental class, the average time needed to finish the posttest is 3.85 minutes, while the average time used by the control group in the posttest is 6.5 minutes. The duration difference indicates the experiment group finished the test questions with greater ease and confidence, which in turn suggests the effectiveness of the self-directed vocabulary learning through COCA.

4.2 The surveys

The results of the questionnaire surveys conducted before and after the tests were found to be in accordance with data analyzed in the above section. About the question concerning subjects' difficulty in producing collocations, 51 participants reported an affirmative answer. This finding concurs with [17]'s view that for L2 learners the process of building syntagmatic connections between words seems to be harder than the process for building paradigmatic connections.

When distinguishing synonymous words, 49 of them said they would turn to dictionaries for help, and 3 subjects would search the internet for other users' answers. Their response shows the dominance of traditional vocabulary learning − dictionary consultation. Students turn to dictionaries for word sense and example sentences illustration when they come to a new word. Dictionaries do fulfill users' needs in explaining word meanings and their usages, yet they fail to provide abundant real contextualized language input. In addition, traditional dictionaries fail to give the collocate profiles for the entries, and collocation dictionaries yet cannot incorporate the contexts of language use. For synonymous words, contexts and real language use are helpful in sense disambiguation. Therefore, participants' judgements on the usefulness of dictionaries were not as positive as the online corpus. Table 7 below presents detailed information.

Table 7. Subjects' response to the two tools tested in this study

	Very helpful	Helpful	Not very helpful	Not helpful	Total
COCA	17	6	3	0	26
Dictionaries	5	6	8	7	26

As is shown in Table 7, most of the subjects in the experimental class (23 out of 26) found it useful to use COCA to acquire vocabulary and thus were in favor of using COCA for vocabulary learning; the data for the control group presented a mixed picture. Most of the subjects in the control class (15 out of 26) did not have very positive attitudes towards using dictionaries in differentiating synonymous words. The remaining 11 students reported that they found dictionaries helpful, but it was worth noting that their scores didn't significantly improve (see Table 6 for statistics). This finding confirmed [16]'s finding that subjects' confidence in performing the collocation task showed the most degree of confidence whilst their score were the lowest when they were given dictionaries for consultation. Such a contradictory situation was not surprising given that traditional vocabulary learning relies heavily on dictionary, or on traditional vocabulary instruction. Teachers firstly provide word meanings and then give students several example sentences, sometimes followed by translation exercises. Without ample language use data, students cannot easily find the nuances of words with similar meanings under traditional vocabulary instruction.

The following discussion takes participants' performance on the synonym pair *fulfill* and *realize* as an example. According to the automatic calculation and analyses of students' scores on the online teaching platform, participants in both groups demonstrated a number of errors with *fulfill* and *realize* in the pretests. Accuracy rates of the

four items were fewer than 50%, which meant a majority of students found it hard to make a distinction between the semantics of the two words. Dictionary definitions (Oxford Dictionary) about *fulfill* and *realize* are (only definitions concerning their mutual sense of 'achieving' are listed):

Fulfill:

1) To do or to achieve what was hoped for or expected: e.g., to fulfill your dream/ambition/potential
2) To do or have what is required or necessary: e.g., to fulfill a duty/an obligation/a promise

Realize: To achieve something important that you very much want to do: e.g., She never realized her ambition of becoming a professional singer.

According to the above definitions, both verbs share common ground in the meaning of 'achieving or making something that is important and valued true', and hence they have an overlap of ranges of collocates, e.g., *ambition, dream*, etc. Based solely on their meanings and collocates or example sentences, it is hard to observe the subtle differences between *fulfill* and *realize*. This pair of near synonyms is distinguished by other range of collocates that are exclusive to only one of them, and these collocates delimit word senses. Therefore, it is not a surprising finding that even participants in the control group felt more confident with the aid of dictionaries, the accuracy rates of the items concerning *fulfill* and *realize* did not show a marked improvement in the posttests. Students failed to fully acquire the semantics of the two verbs through traditional dictionary consultation. On the contrary, the accuracy rates improved markedly in the experimental group (from 40% in general to nearly 80%). Such an improvement indicated that subjects noticed more nuances embedded in contexts under the data-driven learning model. Concordance tools and computational functions provided by COCA enable students to generalize the usage patterns and fully understand word semantics. *Fulfill* places more emphasis on the sense of 'carry out' whist *realize* stresses more on 'reach / make something true'. So, the four sentences in the tests are not difficult to complete:

1. It was really disheartening that our worst fears were_____.
2. They _____ all the boss' demands just to please him.
3. The technicians are trying out new methods _____their design possibilities.
4. The company has taken all measures to _____ the contract by the end of the year.

Sentences 1 and 3 involve the sense of 'reach/make true' whilst 2 and 4 refer to the 'carrying out' of demands and contract. As is generalized from the online corpus in Table 5 in Section 3, *fulfill* is followed by *promise, need, obligation, dream, requirement, role, duty, and realize takes potential, dream, ambition* as it collocates. With the above information, students can easily choose the correct verbs among the near synonym pairs.

5 Discussion

The present study set out to design a data-driven vocabulary learning model based on COCA, and to investigate whether the model boosts effectiveness compared with traditional vocabulary learning. Research findings are:

1. There was an improvement in the acquisition of the tested vocabulary, either with the data-driven learning method or dictionary-consultation method. However, the experimental group under the DDL model produced higher mean scores and had a more reliable performance (as the standard deviation showed). A T test showed a significant relationship between the pretest scores and posttest scores among the experimental group whereas the scores in the control group did not reach statistical significance. That suggested a significant improvement in learners' performance on collocation production under the data-driven learning model, whilst dictionaries did not prove to have such a contributing effect.
2. From the perspective of the time needed for the posttests, the experimental group submitted the tests within a much shorter time and showed more confidence and ease.
3. In terms of attitude factors, nearly 89% of the subjects in the experimental class found it was useful to use COCA and would like to use it in the future for vocabulary learning.

Therefore, it can be concluded that learners' collocation performances have been improved through using the data-driven learning model. Even though new technologies in identifying and correcting learners' misuse of collocations emerge in great quantity, they are not helpful for learners to internalize word semantics. For instance, although EFL writing software like AccurIT is found helpful in improving students' ability to write correct collocations [14], the ultimate aim for enhancing idiomaticity without any aids cannot be attained. Instead, data-driven learning can on the one hand improve students' independent learning ability; on the other hand, it contributes to internalized learning, by which knowledge can be retained and retrieved when there is need for communication.

6 Conclusion and Recommendations

As the building block of a language, vocabulary acquisition has long been on the center stage of second language learning. It also constitutes a difficult domain in the production process for L2 learners. Synonymous word pairs and their collocations baffle even the most proficient learners. The past decades have seen a dramatic increase in studies on synonym distinction and collocation production. Data-driven learning, with the aid of modern computer science and technology, provides a viable and practical option for vocabulary learning, and hence for nativelike idiomatic language production. Through presenting students how the online corpus – COCA can contribute to the differentiation of synonyms; this study showed the usefulness of data-driven learning in meaning differentiation and in vocabulary acquisition. On a

macro level, the interface of COCA presents which genres the target words for investigation are most frequently used, their collocates and clusters; on a micro level, the online corpus enables learners to look through the concordances of the search words for a closer look at their real language use. In this sense, the data-driven learning model proposed in this study is not only helpful for the semantic feature identification on the part of learners, but also is beneficial for exploratory learning.

Given the finding that the resources and functions provided by COCA can benefit learners' vocabulary acquisition, the circumstances in which the results were obtained should be further taken into account. First, participants in this study were English majors at the end of their first year in university, so they have accumulated linguistic knowledge and abilities sufficient for generalizing rules by themselves. Learners at lower proficiency level might need more teacher guidance and instruction as to how to use COCA and how to read concordances. There lies a practical need for a systematic training scheme developed for learners who have not developed competence in applying online learning technologies and in generalizing language rules. Second, the numbers of subjects (26) and the vocabulary tested (5 pairs of verbs) were modest, which means in the future more students are to be included for verification. Third, the experimental group was given a short period of only two classroom hours (90 minutes) for independent learning (teacher's demo included), so a longer span of learning of an expanded list of words can be covered in future study. In sum, a varied profile of words shall be incorporated in the data-driven learning model, and participants of different proficiency levels shall be included in future empirical studies.

7 References

[1] Wilkins, D. A. (1972). Linguistics in language teaching. Cambridge: MFT Press.
[2] Soruç, A., Tekin, B. (2017). Vocabulary learning through data-driven learning in an English as a second language setting. Educational Sciences: Theory & Practice, 17(6), 1811–1832. https://doi.org/10.12738/estp.2017.6.0305
[3] Flowerdew, L. (2012). Corpora and language education. New York: Palgrave Macmillan.
[4] Sinclair, J. (2004). Trust the text. London: Routledge.
[5] Wray, A. (2002). Formulaic language and the lexicon. Cambridge: Cambridge University Press.
[6] González Fernández, B., Schmitt, N. (2015). How much collocation knowledge do L2 learners have? The effects of frequency and amount of exposure. International Journal of Applied Linguistics, 166(1), 94-126. https://doi.org/10.1075/Itl.166.1.03fer
[7] Schmitt, N., Carter, R. (2004). Formulaic sequences in action: An introduction. In: N. Schmitt, ed. Formulaic Sequences: Acquisition, Processing and Use. Amsterdam: Benjamins: 1-22. https://doi.org/10.1075/lllt.9.02sch
[8] Laufer, B., Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. Language Learning, 61(2), 647-672. https://doi.org/10.1111/j.1467-9922.2010.00621.x
[9] Men, H. (2018). Vocabulary increase and collocation learning: A corpus based cross-sectional study of Chinese learners of English. Singapore: Springer Nature. https://doi.org/10.22168/2237-6321-21472

[10] Jiang, N. (2002). Form-meaning mapping in vocabulary acquisition in a second language. Studies in Second Language Acquisition, 24, 617-637. https://doi.org/10.1017/S0272263102004047
[11] Harvey, K., Yuill, D. (1994). The COBUILD testing initiative: The introspective, encoding component. Unpublished Research Paper.
[12] Carter, R., McCarthy, M. (1988). Vocabulary and language teaching. London: Longman.
[13] Wang, X. F., Davies, M., Liu, G. H. (2008). A good platform for English teachers and learners: The corpus of contemporary American English (COCA). Technology Enhanced Foreign Language Education, 123, 27-33.
[14] Grami, G. (2020). An evaluation of online and automated English writing assistants: Collocations and idioms checkers. International Journal of Emerging Technologies in Learning (iJET), 15(4), 218-226. https://doi.org/10.3991/ijet.v15i04.11782
[15] Firth, J. R. (1957). Papers in linguistics 1934-1951. London: Oxford University Press.
[16] Lee, C. Y., Liu, J. S. (2009). Effects of collocation information on learning lexical semantics for near synonym distinction. Computational Linguistics and Chinese Language Processing, 14(2), 205-220.
[17] Wolter, B. (2006). Lexical network structures and L2 vocabulary acquisition: The role of L1 lexical/conceptual knowledge. Applied Linguistics, 27(4), 741-747. https://doi.org/10.1093/applin/aml036

8 Author

Haiyan Men is a lecturer in the Faculty of Foreign Languages of Shanghai Sanda University, Shanghai 201209, China. She graduated in 2015 from Birmingham City University in the UK with a PhD degree in corpus linguistics and applied linguistics.

Article submitted 2020-10-03. Resubmitted 2020-11-05. Final acceptance 2020-11-11. Final version published as submitted by the authors.