Practice-Oriented Paper

Using Data-Driven Learning to Discover Patterns of Grammar

Sarah Miyoshi Deutchman
Waseda University

Data-driven learning (DDL) is a powerful tool that can help students find patterns in language on their own. Unlike a teacher or professor, corpora are available to students whenever they wish to consult them. In fact, students can feel empowered by using corpora. Additionally, corpora can help to clarify notions found in a language. However, corpus consultation is not something that is often taught to learners. This might be because using corpora to search for patterns in language can appear intimidating at first. More user-friendly websites such as Sketch Engine for Language Learning (SKELL) and Contemporary Corpus of American English (COCA) have made corpus consultation more accessible to students. With some training, students can gain information that will help them improve their language skills. This paper gives an overview of grammatical patterns found in corpora, discusses how to do basic searches, and suggests some activities that can be used in class.

Corpora can be defined as "a collection of naturally occurring, computer-readable texts, often comprising many millions of words, which is considered more or less representative of a particular domain of language use" (Hyland, 2013, p. 248). Although corpora can arguably be made from any source of information from English for Specific Purposes (ESP) to academic writing, it is necessary to choose corpora based on the needs of the course. Using corpora allows students to consult with language as it naturally exists, which is important because sometimes language used in textbooks and other resources may not be used in a more interactive context (Coxhead et al., 2017). Thus, if a student encounters words outside of class, there is a possibility that they may not recognize it as a word that they already know. This is because different areas of study can use highly frequent words in a specialized way. However, it may take a considerable amount of time

to create and compile one's own corpora, which could be too complicated for students or some teachers who are new to corpus linguistics.

However, it is possible to find ready-made corpora online available for free, such as Corpus of Contemporary American English (COCA) or Sketch Engine for Language Learning (SKELL). Even though these corpora allow students to interact with authentic language, that interaction is pointless if students cannot comprehend what they are reading. Thus, it is important to decide which corpus to choose depending on students' levels. It is also necessary to consider how and where the corpus is going to be used (Charles, 2018). COCA allows users to look up genre, keyword in context (KWIC), compare words, see collocates, and look for historical patterns (Davies, 2020). COCA also has a word function that displays many of these features on one page. Because of the abundance of information that can be found on COCA, it may take a bit more training to learn how to use it. On the other hand, SKELL is a bit more user-friendly in that it only requires users to type in a single word. It, like COCA, can display the word in a sentence (i.e., concordance lines). However, unlike COCA, the lines on SKELL are not truncated. Hirata and Hirata (2019) found that clean concordance lines do not overwhelm learners with too much information and are more user-friendly. Additionally, it is possible to use the word sketch function to look for collocates (i.e., words associated with each other) and see how those words are used together in a sentence. The downside of using SKELL is that it works best with more frequently occurring vocabulary. For a word such as *uxorious*, only the concordance lines will appear.

Students can analyze corpora using data-driven learning (DDL), which is an approach that uses computers as an intermediary to help students discover language patterns on their own (Johns, 1991); it works well for learning vocabulary items, basic grammar items, and verb phrases (Mizumoto & Chujo, 2015). Because DDL relies on users to notice the patterns within the corpora, it requires teachers to act as directors to aid students in their research (Johns, 1991). One of the main tools used in DDL is a concordancer, which is a tool that shows how the word is used in a variety of contexts in various forms that can be easily searched (Johns, 1991). However, students will likely need guidance

when using a concordancer for the first time. Therefore, the teacher needs to choose a question to find the appropriate information and then model the action for the students (Cobb & Boulton, 2015). This paper aims to make corpora and DDL more accessible to those who have not had experience with either before by explaining the basic terminology used in DDL, talking about how to do basic searches, and discussing how to plan activities that can be used in the classroom.

Grammatical Patterns Found in Corpora

DDL gives students the ability to interact with authentic language which makes them more sensitive to its use and increases their metacognitive skills (Boutlon & Tyne, 2013). Moreover, DDL offers a flood of information that can be used to help learners deal with the fuzzy nature of authentic language (Boulton, & Cobb, 2017). Boulton and Pereiro (2008) have stated that neither traditional reference sources nor intuition can helpfully disambiguate two similar words (e.g., *almost* and *nearly*). Therefore, using corpus consultation is beneficial. However, learners need to be able to notice the patterns (Schmidt, 1990 as cited in Boulton & Cobb, 2017) found in corpora such as collocations, colligations, and lexical bundles.

When studying new words, it is important to understand which words usually occur together (i.e., collocations). This is because understanding collocations is part of knowing a word (Schmitt, 2010). The classic example of collocation use is *strong coffee* versus *powerful coffee*. Although there is nothing grammatically wrong with *powerful coffee*, the phrase sounds unnatural to a native speaker. This is because knowledge of when and where to use collocations may be different between native speakers and non-native speakers of a language (Zhang, 1993). An example of this is the word *wear* in English. In Japanese *wear* can be translated to *kiru, kaburu, haku, hameru, kakeru, shiteru,* and *chakuyou shiteiru* depending on what part of the body is being talked about. If one of the Japanese verbs for wear was matched with the wrong body part, the writing would appear unnatural, which could lead to a lower quality of writing. On the other hand, those who have a more accurate knowledge of collocations have a higher quality of writing (Zhang, 1993). Therefore, teaching collocation use is important.

Another grammatical pattern that can be investigated by using DDL are

colligations, environments in which a word is used, what words are connected to it, and what the grammatical patterns are (Sinclair, 1996, 1999, 2004 and Hoey, 1997a, 1997b, as cited in Hoey & O'Donnell, 2008). For example, despite + det + adj + noun (e.g., Despite a serious injury; Despite a pristine record; Despite a decent cast). Colligations can confuse students if they are not aware of which pattern goes with which word. An illustration of this is understanding how and when words should be used in context. *While* is usually used with a verb + ing form (e.g., while she was dreaming). *During* is used with a period of time (e.g., during the 1980s, during World War 2). The difference between using *while* and *during* can become clearer when consulting corpora.

The final language pattern in this section is lexical bundles which can be explained as extended collocations that frequently appear together (Biber et al., 1999, as cited in Hyland, 2012). It is important to study lexical bundles as they can aid in writing proficiency. Understanding how these multi-word units are used helps improve fluency (Hyland, 2012). By using the KWIC on COCA it is possible to search for strings, such as *due to the fact* where it can be seen that *due to the fact* is more likely a reduced form of *due to the fact that*. This result shows students what the full lexical bundle looks like and how to use it in context.

Basic Searches in COCA

The section addresses how to do some basic searches in COCA, such as collocations, comparison, KWIC, and genre. When using the collocation function, it is possible to adjust for how many words in front of or behind a word should be examined. Thus, if one were to look at which adjective were connected to a noun, it would make sense to look at one or two words in front of that noun and select adj for the collocate (Figure 1). Results for a typical search on the word culture show the top ten most common collocates in order (Figure 2).

Comparison searches can show which nouns should go with *much* or *many*. To do this search it is necessary to select nouns from the POS options and limit them to one word behind *much* or *many*. This is done because nouns would occur one space behind *much* and *many* in a sentence (Figures 3 and 4).

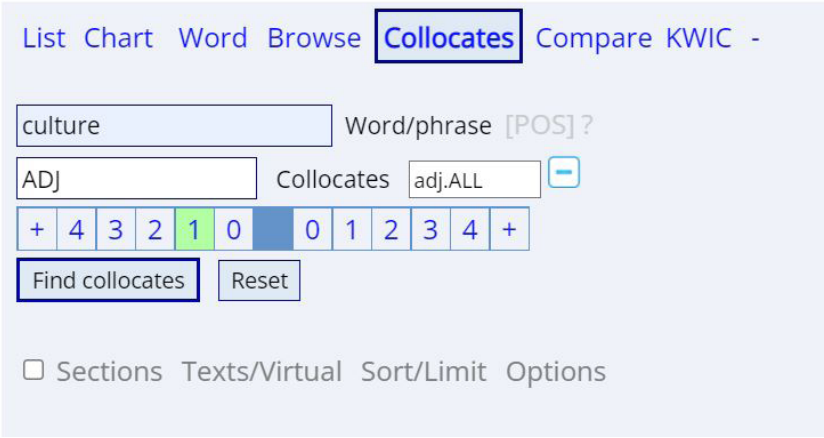


Figure 1. How to do an adjective collocate search *culture* + *adj*.
Note. It is necessary to push the (+) button to see the collocate search option. The POS can be changed depending on what pattern is being examined.

HELP			FREQ	
1	☐	POPULAR	3613	
2	☐	AMERICAN	3524	
3	☐	POLITICAL	1251	
4	☐	WESTERN	1195	
5	☐	CORPORATE	743	
6	☐	DOMINANT	668	
7	☐	DIGITAL	551	
8	☐	CHINESE	515	
9	☐	MASS	503	
10	☐	BLACK	475	

Figure 2. Results of the collocation search.
Note. It can be seen that *popular* is most commonly associated with *culture*.

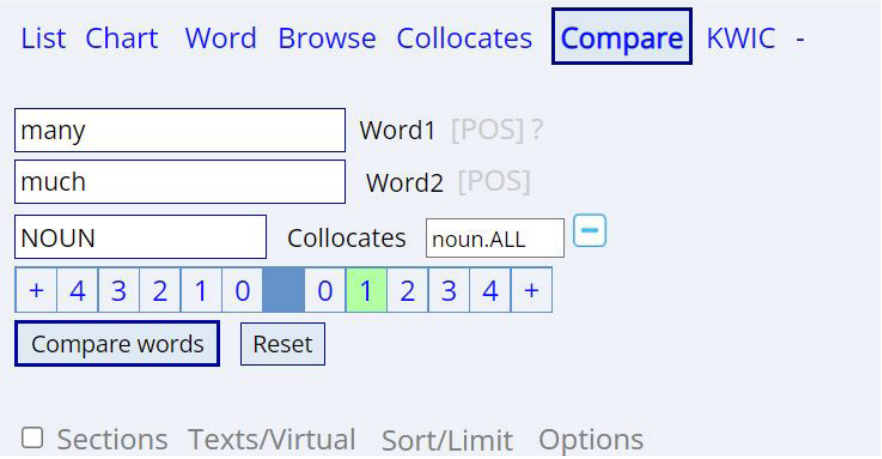


Figure 3. How to do a comparison search with *much* and *many*.

WORD 1 (W1): **MANY** (0.94)

	WORD	W1	W2	W1/W2	SCORE
1	WAYS	14241	0	28,482.0	30,454.4
2	YEARS	26516	2	13,258.0	14,176.1
3	CASES	8222	1	8,222.0	8,791.4
4	AREAS	2169	0	4,338.0	4,638.4
5	INSTANCES	1493	0	2,986.0	3,192.8
6	RESPECTS	1470	0	2,940.0	3,143.6
7	REASONS	2923	1	2,923.0	3,125.4
8	PLACES	2904	1	2,904.0	3,105.1
9	TIMES	28337	10	2,833.7	3,029.9
10	FORMS	1413	0	2,826.0	3,021.7

WORD 2 (W2): **MUCH** (1.07)

	WORD	W2	W1	W2/W1	SCORE
1	FUN	5695	0	11,390.0	10,652.3
2	HELP	742	0	1,484.0	1,387.9
3	EMPHASIS	688	0	1,376.0	1,286.9
4	CHOICE	609	0	1,218.0	1,139.1
5	EFFORT	1176	1	1,176.0	1,099.8
6	ATTENTION	4274	4	1,068.5	999.3
7	SUPPORT	518	0	1,036.0	968.9
8	RIGHT	512	0	1,024.0	957.7
9	INFLUENCE	506	0	1,012.0	946.5
10	ADO	463	0	926.0	866.0

Figure 4. Results of the comparison search of *much* and *many*
Note. Countable nouns occur with the word many (only countable nouns use the plural -s) and uncountable nouns occur with the word much.

By using these basic searches, it is possible to create a multitude of DDL-based activities that can be done by students.

KWIC can be used to show a single word or phrase together in context (Figures 5 and 6). It is possible to set options to choose different contexts (e.g., academic, blog, news, spoken).

Genre searches can be used to assist with academic writing as students will be

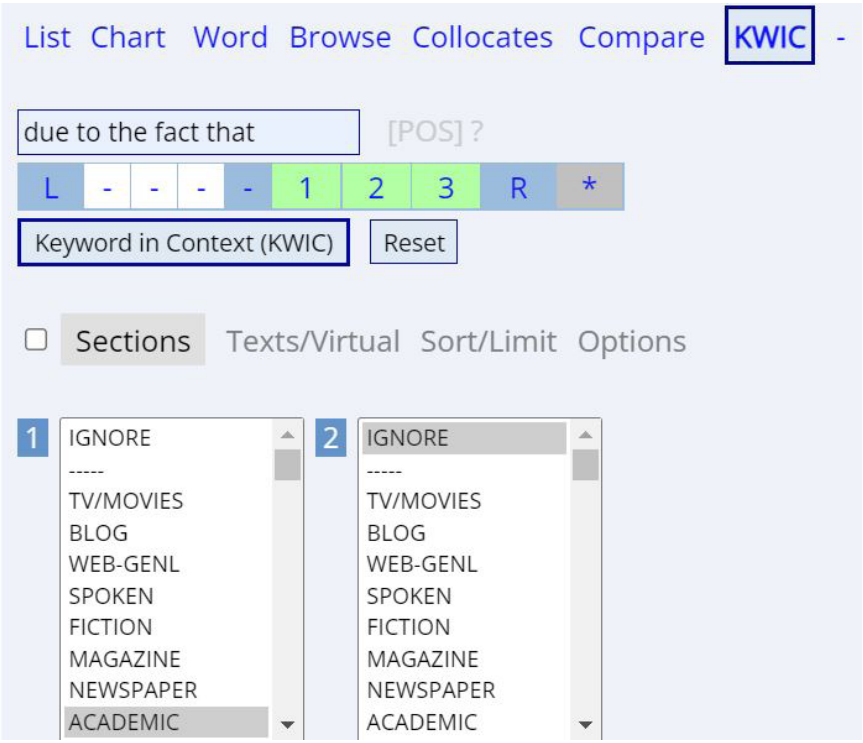
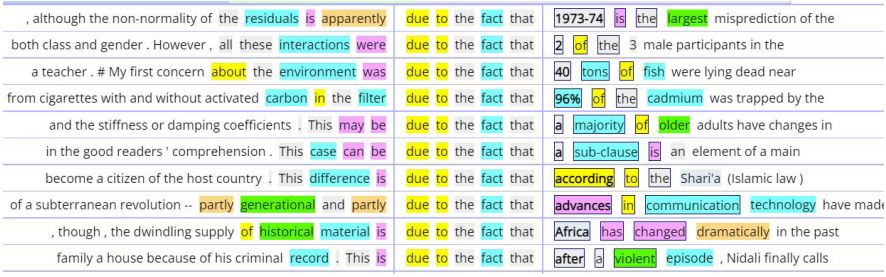


Figure 5. KWIC search with 'Due to the Fact That' limited to academic writing.



, although the non-normality of the residuals is apparently	due to the fact that	1973-74 is the largest misprediction of the
both class and gender . However , all these interactions were	due to the fact that	2 of the 3 male participants in the
a teacher . # My first concern about the environment was	due to the fact that	40 tons of fish were lying dead near
from cigarettes with and without activated carbon in the filter	due to the fact that	96% of the cadmium was trapped by the
and the stiffness or damping coefficients . This may be	due to the fact that	a majority of older adults have changes in
in the good readers ' comprehension . This case can be	due to the fact that	a sub-clause is an element of a main
become a citizen of the host country . This difference is	due to the fact that	according to the Shari'a (Islamic law)
of a subterranean revolution -- partly generational and partly	due to the fact that	advances in communication technology have made
, though , the dwindling supply of historical material is	due to the fact that	Africa has changed dramatically in the past
family a house because of his criminal record . This is	due to the fact that	after a violent episode , Nidali finally calls

Figure 6. KWIC search with 'due to the fact that' limited to academic writing.

able to see if a word is academic or not (Figures 7 and 8). Students can also check if a word is still being used.

DDL Based Activities

There are many activities learners can do that involve corpora from less planned activities to more highly controlled activities (Boulton & Tyne, 2013). Multiple-

Figure 7. 'A lot of' genre search.
Note. By using the chart function, it is possible to see where and when a word is used.

CHANGE TO VERTICAL CHART / CLICK TO SEE CONTEXT

SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD
FREQ	334000	49109	34726	41136	120423	15808	28769	38868	5161
WORDS (M)	993	20.7	20.1	20.7	20.4	19.1	20.4	19.6	19.3
PER MIL	336.35	2,359.26	1,726.82	1,984.54	5,898.92	825.49	1,409.75	1,972.66	266.20
SEE ALL SUB-SECTIONS AT ONCE									

Figure 8. Results of 'a lot of' genre search.
Note. A lot of would usually be used when speaking and would rarely be used in academic writing.

choice activities are an example of more controlled activities. With this activity, students can be given idioms on one side and definitions of the idioms on the other side. By reading the sentences on COCA, students can get the information to match the idiom to the appropriate definitions (Figure 9). Additionally, it is possible to use coded feedback and have students look at corpus data to fix their errors (Tono et al., 2014). Teachers can look through students' essays and notice common patterns of error. For example, the chart function can be used to show students which words would be used in academic writing.

Although there is no real limit to the number of activities that can be created using DDL, some practices should be used when planning these activities. First, it is best to run the searches and have results printed up in case students' results are not consistent. Corpora can be updated regularly and as a result, search results can change. Second, for beginners, it might be best to start with screenshots of the searches and the results so students can use them for reference. Third, it is important to consider why these tasks are being done (e.g., targeting previous mistakes, showing academic vocabulary). Fourth, time constraints should be considered. It is unlikely that students will have experience with corpora. Therefore, it is necessary to consider how much time is needed to train students. If the materials provided are relatively simple, such as using a gap-fill worksheet with pre-printed search results (Figure 10), students should be able

Find the meaning of the idioms. Look at the concordance lines. What do you think the idioms mean?

1)In light of	a) someone or something that has the power to make something happen
2)Driving force	b) something done if nothing else works
3) Along the lines of	c) because of, considering
4) Last resort	d) similar, alike

Figure 9. Example of a meaning-focused activity.
Note. In this activity, students can use the concordance lines to try to determine the meaning of each idiom.

WORD 1 (W1): **MUCH** (1.07)

	WORD	W1	W2	W1/W2	SCORE
1	FASTER	2435	0	4,870.0	4,554.6
2	ELSE	1976	0	3,952.0	3,696.0
3	EASIER	6788	2	3,394.0	3,174.2
4	LONGER	7730	3	2,576.7	2,409.8
5	DEEPER	1190	0	2,380.0	2,225.9
6	EFFORT	1173	0	2,346.0	2,194.1
7	EVERYTHING	872	0	1,744.0	1,631.0
8	FARTHER	864	0	1,728.0	1,616.1
9	WIDER	767	0	1,534.0	1,434.6
10	HARDER	3011	2	1,505.5	1,408.0

WORD 2 (W2): **MANY** (0.94)

	WORD	W2	W1	W2/W1	SCORE
1	WAYS	14244	0	28,488.0	30,460.8
2	YEARS	26522	2	13,261.0	14,179.3
3	CASES	8223	1	8,223.0	8,792.4
4	REASONS	2927	0	5,854.0	6,259.4
5	AREAS	2167	0	4,334.0	4,634.1
6	INSTANCES	1492	0	2,984.0	3,190.6
7	THOUSANDS	1423	0	2,846.0	3,043.1
8	TIMES	28347	10	2,834.7	3,031.0
9	FORMS	1413	0	2,826.0	3,021.7
10	LEVELS	1366	0	2,732.0	2,921.2

1) This activity is (much / many) harder than I thought.
2) Surveillance should not be allowed for (much / many) reasons.
3) It took (much/many) times to get it right.
4) The police put (much/ many) effort into catching the criminal.
5) What do you notice about the words used with much and the words used with

Figure 10. Example of a gap fill activity.

to do the worksheet independently after the activity is modeled for them. With more complex, less controlled activities it can take a couple of hours of training. Fifth, and perhaps the most important, is to consider the students' language proficiency level. DDL-based activities are something that should not be painful to experience. If students are at a lower level, it is possible to use bilingual corpora and have students compare patterns between their L1 and L2. Sixth, it is important to remember that learning to consult corpora is a process. Even if students have a lesson on using corpora it does not automatically mean there will not be any errors in their writing. Lee et al. (2009) found that the writing products of their students still had some errors which might have been due to incorrect concordancing, issues during training, or different ideas on grammar. Thus, even though DDL cannot solve all writing problems, it is still helpful.

Conclusion

Although using corpora and DDL may feel daunting at first, with some training both teachers and students can use them effectively. DDL allows students to gain autonomy by being able to use corpora whenever they desire. Additionally, students can understand which words are usually associated with each other. In the future, DDL should be built into curriculums as it helps learners reach a

higher language proficiency.

References

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, *67*(2), 348–393. https://doi.org/10.1111/lang.12224

Boulton, A., & Pereiro, M. (2008). But what's wrong with it? Corpus linguistics, helping non-linguists find order in a fuzzy world. In M. Pereiro & H. Daniels (Eds.), *Le désordre. Grendel, n° special* (pp.161–185). AMAES. https://hal.archives-ouvertes.fr/hal-00327220/

Boulton, A., & Tyne, H. (2013). Corpus linguistics and data-driven learning: A critical overview. *Bulletin Suisse de Linguistique Appliquée*, *97*, 97–118.

Charles, M. (2018). Corpus tools for writing students. *The TESOL Encyclopedia of English Language Teaching*, 1–7. https://doi.org/10.1002/9781118784235.eelt0554

Cobb, T. & Boulton, A. (2015). Classroom applications of corpus analysis. In D. Biber & R. Reppen (Eds.), *Cambridge Handbook of English Corpus Linguistics* (pp. 478–497). Cambridge University Press. https://doi.org/10.1017/CBO9781139764377.027

Coxhead, A., Dang, T. N. Y., & Mukai, S. (2017). Single and multi-word unit vocabulary in university tutorials and laboratories: Evidence from corpora and textbooks. *Journal of English for Academic Purposes*, *30*, 66–78. https://doi.org/10.1016/j.jeap.2017.11.001

Davies, M. (2020). The COCA corpus. *The Corpus of Contemporary American English*. https://www.english-corpora.org/coca/help/coca2020_overview.pdf

Hirata, Y., & Hirata, Y. (2019). Applying 'Sketch Engine for Language Learning' in the Japanese English classroom. *Journal of Computing in Higher Education*, *31*(2), 233–248. http://dx.doi.org/10.1007/s12528-019-09208-z

Hoey, M., & O'Donnell, M. B. (2008). Lexicography, grammar, and textual position. *International Journal of Lexicography, 21*(3), 293–309. https://

doi.org/10.1093/ijl/ecn025

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics*, *32*, 150–169. https://doi.org/10.1017/S0267190512000037

Hyland, K. (2013). Corpora and innovation in English language education. In Hyland, K & Wong, L. (Eds.) *Innovation and change in language education (*pp. 248–262). Routledge.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *English Language Research Journal, 4,* 1–6.

Lee, M., Shin, D., & Chon, Y.V. (2009). Online corpus consultation in L2 writing for in-service teachers of English. English Teaching, *64*(2), 233–254.

Mizumoto, A., & Chujo, K. (2015). A meta-analysis of data-driven learning approach in the Japanese EFL classroom. *English Corpus Studies, 22*, 1–18.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* Springer.

Tono, Y., Satake, Y., & Miura, A. (2014). The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL, 26*(2), 147–162. https://doi.org/10.1017/S095834401400007X

Zhang, X. (1993). *English collocations and their effect on the writing of native and non-native college freshmen* [Unpublished doctoral dissertation], Indiana University of Pennsylvania.

Author Bio

Sarah Miyoshi Deutchman *teaches at Waseda University as a part-time lecturer. She has taught English for over 10 years in three different countries: America, South Korea, and Japan. Her areas of research include data-driven learning, corpus linguistics, and vocabulary. sarah@aoni.waseda.jp*

Received: October 31, 2021
Accepted: September 27, 2022