Japanese-English Parallel Corpora in the Classroom : Applications and Challenges

関西外国語大学 研究論集 第107号（2018年 3 月）
Journal of Inquiry and Research, No.107 (March 2018)

Japanese-English Parallel Corpora in the Classroom: Applications and Challenges

Michael P. McGuire

Abstract

Computerized corpora have given linguists crucial new insights on the usage of language. With the help of software, it is possible to index the words which appear in a large collection of text and analyze word usage and frequency. Data Driven Learning looks at how students can benefit from their own direct use of corpora. While monolingual corpora have a steep learning curve and are often too difficult for language learners, a solution to this problem may be found in bilingual parallel corpora, which are built from authentically translated text. This article looks at Eijiro on the WEB and Weblio, two online Japanese-English parallel corpus based websites. Some guided practice exercises developed by the author for use in university level English language writing classes in Japan are discussed, and some of the challenges in training students to use these resources to improve their English language writing are presented.

Keywords: corpus linguistics, parallel corpora, data driven learning, collocation, English language education in Japan

Background

Computerized corpora have given linguists crucial new insights on the usage of language. With the help of software, it is possible to examine a collection of text and index the words which appear in it (known as concordancing). By compiling large collections of texts taken from authentic sources like magazines, newspapers, and books, researchers can analyze word usage and frequency. Numerous online corpora have been created and are open to the public. One well known example is the Corpus of Contemporary American English (Davies, 2008), or COCA, which contains over 520 million words from a wide range of genres and contexts (spoken, written, academic, news, fiction, etc…).

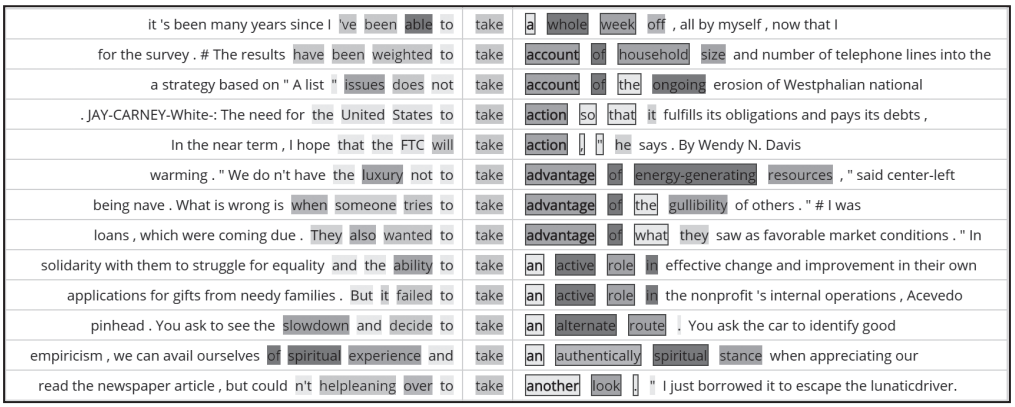
By concordancing large samples of real language, more authentic resources based on statistical data can be developed for language education. The ability to analyze word frequency and usage in authentic texts can help in the creation of such materials (Conrad,

2000). Recently, many new resources are corpus informed. The Collins Cobuild series of learner dictionaries all include example sentences taken from The Bank of English corpus. Cambridge's Touchstone series of textbooks are designed around authentic word usage pulled from the Cambridge International Corpus. The Academic Vocabulary List, or AVL, compiles the top 3000 lemmas (the dictionary form of a word) found in academic texts in COCA.

Additionally, many researchers have developed learner corpora (see the Learner Corpora Association website). These corpora are collections of texts produced by non-native learners. Some are focused on learners from a specific first language, while others compile large varieties of languages. By closely examining the second language production of learners, researchers can identify common mistakes and patterns, which can inform creation of textbooks and other teaching materials.

Materials development can clearly benefit from corpus linguistics, but Data Driven Learning, or DDL (Johns & King, 1991), looks at how students can benefit from their own direct use of corpora. It is easy to see the potential benefits: a corpus provides extensive authentic context for lexical items. Language students commonly look for meanings of unfamiliar words in bilingual dictionaries, but the translations presented are decontextualized, which makes it difficult for students to select the appropriate one. While many newer dictionaries provide a sample sentence or two, this is not enough for students to be certain of the correct usage. But because a corpus is composed exclusively of authentic text, a student could search a corpus for a word or phrase and see a large variety of examples of real world usage. With some effort, a student can examine enough sentences to find the appropriate meaning and usage of a word.

Collocations are words which frequently appear together and may have a more powerful lexical function that their constituent parts, such as "heavy smoker" or "take a risk". These are usually not included in dictionaries, but corpus software can easily identify them. Many corpora can display search results in the Key Word in Context (KWIC) format, which will align the search term down the center of the page, making it easy to see the words that appear in proximity. Figure 1 shows KWIC results for the search term "take" in COCA. The words "account", "action", and "advantage" immediately follow in multiple sentences, suggesting that these combinations are common.



it 's been many years since I 've been able to	take	a whole week off , all by myself , now that I
for the survey . # The results have been weighted to	take	account of household size and number of telephone lines into the
a strategy based on " A list " issues does not	take	account of the ongoing erosion of Westphalian national
. JAY-CARNEY-White-: The need for the United States to	take	action so that it fulfills its obligations and pays its debts ,
In the near term , I hope that the FTC will	take	action , " he says . By Wendy N. Davis
warming . " We do n't have the luxury not to	take	advantage of energy-generating resources , " said center-left
being nave . What is wrong is when someone tries to	take	advantage of the gullibility of others . " # I was
loans , which were coming due . They also wanted to	take	advantage of what they saw as favorable market conditions . " In
solidarity with them to struggle for equality and the ability to	take	an active role in effective change and improvement in their own
applications for gifts from needy families . But it failed to	take	an active role in the nonprofit 's internal operations , Acevedo
pinhead . You ask to see the slowdown and decide to	take	an alternate route . You ask the car to identify good
empiricism , we can avail ourselves of spiritual experience and	take	an authentically spiritual stance when appreciating our
read the newspaper article , but could n't helpleaning over to	take	another look . " I just borrowed it to escape the lunaticdriver.

Figure 1. An example of KWIC format for the search term "take" in COCA.

By taking this a step further, we can rank collocations based on the number of occurrences in the corpus. Figure 2 shows a ranked list of words that appear immediately after the search term "take". The word "care" appears most frequently in this position with 15371 occurrences in the corpus, followed by "place" and "advantage". Words in a collocation do not always appear side by side, so it is also possible to specify different position ranges, such as words that appear within four positions before or after the search term. This search display format as well as the KWIC display format can potentially help students with finding collocations, determining article usage, and identifying appropriate prepositions before and after words.

		CONTEXT	FREQ	ALL	%	MI
1	☐	CARE	15371	166639	9.22	7.01
2	☐	PLACE	7682	251124	3.06	5.41
3	☐	ADVANTAGE	6328	30104	21.02	8.19
4	☐	YOUR	6293	884008	0.71	3.31
5	☐	OVER	6041	671424	0.90	3.65
6	☐	THEM	5624	888475	0.63	3.14
7	☐	OFF	4969	382410	1.30	4.18
8	☐	AWAY	3205	259256	1.24	4.11
9	☐	PART	2465	259361	0.95	3.73

Figure 2. Ranked frequency of words that appear immediately after the search term "take" in COCA

It is clear that corpora can be powerful tools for language study. However, they have a steep learning curve and are often too difficult for language learners (St John, 2001; Kosem,

2008). Several studies have looked at students' use of corpora to improve their writing skills and correct errors. They found that while students are very receptive to using corpora, they often struggle to comprehend search results (Yoon & Hirvela, 2004; Hegelheimer, 2006; Yoon, 2008; Yoon & Jo, 2014; Luo & Liao, 2015; Luo 2016). The corpora used in these studies were in English only, which could account for some of the difficulty students faced. A solution to this problem may be found in bilingual parallel corpora.

Parallel Corpora

Parallel corpora are built from authentically translated text, such as bilingual publications or translated Wikipedia articles. Concordance searches can be done in a parallel corpus in either language and will produce results in both languages. Therefore, these corpora can be useful even to beginner level students (St. John, 2001). With a Japanese-English parallel corpus, for example, a user can search for words in either language and immediately see all instances of their search term with matching translations. This opens up many possibilities for applications in language learning programs and direct student use.

Chujo, Utiyama, and Miura (2006) looked at the potential for Japanese students to use a Japanese-English parallel corpus for building vocabulary. Their study found that while it took some time to get used to using this new tool, the students found it quite useful, especially for learning collocations. In addition, many of the students stated that the large variety of sentences and meanings found in the corpus helped to raise their awareness of the many possible translations, or the one-to many relationship, between English and Japanese.

There are two websites built upon Japanese-English parallel corpora that can easily be used by students and instructors: *Eijiro on the WEB*, and *Weblio*. Both of these sites are free and are immediately accessible to English and Japanese learners alike.

Eijiro on the WEB (alc.co.jp)

Eijiro (英辞郎) is an English-Japanese translation corpus started in 1998 by the Electronic Dictionary Project. It is regularly updated and presently contains over 2,050,000 lemmas in both English and Japanese from more than 702,000 professionally translated texts, such as articles from bilingual publications such as Asahi Weekly and Hiragana Times.

The most popular online search engine for the *Eijiro* corpus is called *Eijiro on the WEB*

(www.alc.co.jp), created in 2000 by ALC, a Japanese publishing company that focuses on English language educational materials. The search engine, seen in Figure 3, is streamlined and does not have as many advanced search options as COCA, but it does have limited lemma, wildcard, and a few other syntax search abilities. Students can use this search engine easily to search in English or Japanese.

Figure 3. ALC's *Eijiro on the WEB* search engine

A basic search for a single word in English provides a variety of information. Figure 4 shows the topmost results for the word "take". At the very top, other forms of "take" can be seen, followed by a list of idioms and phrases that contain the word "take". All of these words and expressions are hyperlinks, which can open a new corresponding search. Just below the list of idioms, we can see single word translations, similar to a dictionary entry, with a few example sentences. Again, any of the English or Japanese words can be searched in a new window by double-clicking them.

Figure 4. Topmost results for "take" in *Eijiro on the WEB*

Following the single word definitions are translations of short phrases that contain the search term. Figure 5 shows some of the results from the first page of a search for the word “take”, presented in alphabetical order of subsequent words beginning with wildcard ~ placeholders. The search term is colored red, and the rest of the words are colored blue, making it very easy to read.

Figure 5. Phrases from the first page of results for the word “take” in *Eijiro on the WEB*

Other valuable results can be found deeper into the list. After short phrases, longer results and full sentences can be found. Figure 6 shows some results from the fourth page of a search for the word “taking”. These longer results show much more context.

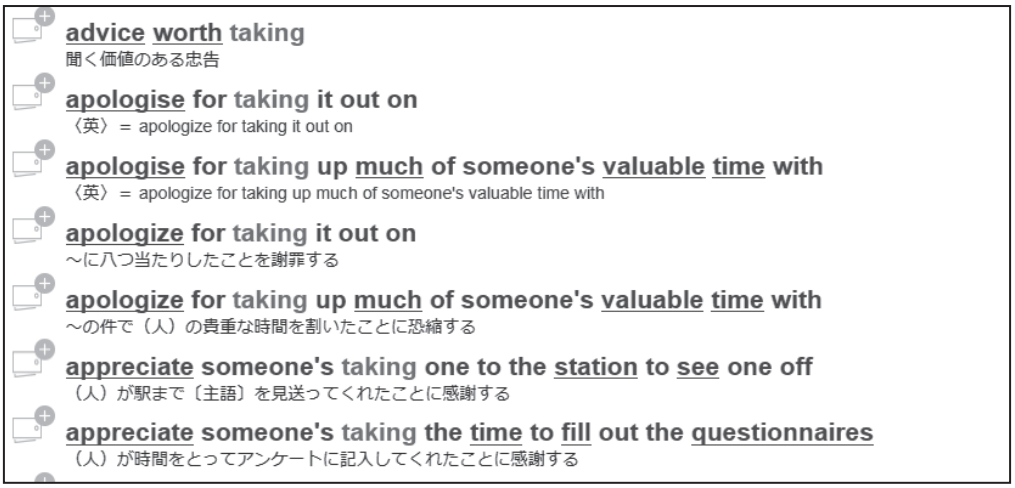


Figure 6. Some results from the fourth page of a search for "taking" in *Eijiro on the WEB*.

Multi-word searches will display results that contain all the words included in the search. Figure 7 shows some of the results from a search for the words "take" and "place". Multi-word search results are sorted alphabetically and by proximity, so results containing the words side by side will appear first, followed by results containing words in between.

Figure 7. Some results from a multi-word search for "take" and "place" in *Eijiro on the WEB*.

Weblio (ejje.weblio.jp)

Weblio is a website with a variety of tools like a bilingual Japanese-English dictionary, machine translation, and vocabulary tests. Though not as powerful as *Eijiro on the WEB, Weblio* also has some very useful corpus-based features. *Eijiro on the WEB* requires a paid subscription to access KWIC functions, but *Weblio* offers them for free. To use the KWIC function, you can click on the concordance tab (共起表現) at the top of the main page, seen in Figure 8, or visit ejje.weblio.jp/concordance/.

Figure 8. The *Weblio* search engine

A single word search will result in basic KWIC results. While they are presented only in English without Japanese translations, the site navigation is in Japanese, making it easier for students to use than a monolingual corpus. Figure 9 shows KWIC results for the word "take".

Figure 9. KWIC results for the word "take" in the concordance (共起表現) section of *Weblio*.

The default view is 1語右で並び替え which highlights the first word after the search term in blue, but more useful results are obtained by clicking on the 隣合う単語の集計結果を見る button to view collocations ranked by frequency. Figure 10 shows collocation frequency results for the word "take". The search term appears in the center column and the columns to the right and left show the words that appear most frequently in positions one or two places before or after the search term. To the right of the search term, we can see that "place" is the most common word that immediately follows "take", appearing 426 times within the corpus of sample sentences collected on the website. Likewise, to the left we can see that "did" appears most frequently two places before "take" with 60 occurrences. Students can use this function of *Weblio* to find collocations, articles, and prepositions that frequently appear in proximity to their search term.


2語左で並び替え　1語左で並び替え　隣合う単語の集計結果を見る　1語右で並び替え　2語右で並び替え

※単語をクリックすると、その単語と隣合っている例文のみ表示します。

2語左の単語		1語左の単語		検索キーワード	1語右の単語		2語右の単語	
did	60	to	1237	take	place	426	in	272
the	57	will	260		the	374	on	170
and	56	and	109		a	271	the	157
eclipse	53	can	98		over	122	of	142
The	51	not	96		part	118	to	115
It	39	would	80		on	105	at	60
decided	36	alternate	67		up	104	from	58
scheduled	33	also	52		Me	67	and	48
I	32	Can't	45		it	64	a	41

Figure 10. Collocation frequency results for "take" in the concordance section (共起表現) of *Weblio*.

How I use these sites in my classes

Here I would like to discuss how I have attempted to use these parallel corpora based websites in my classes. While Data Driven Learning is a broad field with many approaches, a common theme throughout is the exploratory nature of corpus use (Boulton 2009, 2010, 2011; Smart 2014). With this in mind, I have always tried to promote *Eijiro on the WEB* as a resource where students can discover language usage through their own active exploration. I believe that promoting active involvement in language study will lead to greater learner

autonomy, which can benefit students in all of their courses. Taking this kind of initiative is not easy for some students, so I have worked to create guided practice activities to help students learn to use these sites effectively. I am still working to improve this training, and in the last section I will discuss the challenges I have encountered and possible reasons for them.

I introduce both *Eijiro on the WEB* and *Weblio* to all my students. Training students to use *Eijiro on the WEB* is much more involved than *Weblio*. Although the KWIC section of *Weblio* is straightforward and easy to use, *Eijiro on the WEB* is not as intuitive. I spend much more time exploring the latter site with my students, especially in my writing classes. My hope is that they can use this site to explore word choice, collocation, and usage when first writing. Then, in the revision stages, they can use it to help with error correction. There is debate as to whether explicit error correction from a teacher is helpful (Truscot 1996; Ferris 1996), but if students can use these resources to correct mistakes on their own, there might be greater retention of the information.

I introduce *Eijiro on the WEB* to my students by explaining what a corpus is and how it is different from a dictionary. On a projector, I do a few searches and show students what the results look like and how to understand and navigate them. Next, I give my students some time to try searching for any words they like. I encourage them to try searches in English and Japanese, and to try single and multi-word searches.

Next, I show students how *Eijiro on the WEB* can be used for error correction with a few examples from student writing. I tell students that they should try to find examples in *Eijiro on the WEB* that are similar to what they want to write in Japanese and then look at how they are translated into English. If they find an entry syntactically similar to their idea, they can modify the translation for their own writing. I begin with a simple example taken from one student's attempt to translate into English a paragraph she wrote in Japanese:

Original Japanese:	また、私は紅茶風味のお菓子も好きです。
Student English Translation:	*Also, I love sweets which taste is black tea.*

In this case, the student was not sure how to translate "紅茶風味のお菓子" (*koucha fuumi no okashi*) into English. I ask the students to search for this entire phrase in Japanese, but there are zero results. This search term is too specific, so it needs to be reduced in order to find syntactically similar results. I ask the students to think about which parts of the expression can easily be changed. After some discussion, we decide that both "紅茶" (*koucha*

– “black tea”) and “お菓子” (*okashi* – “sweets”/“candy”) can be replaced with other words. This leaves us with “風味” (*fuumi* – “flavor”/“taste”), which is the core of the translation error. Next, I ask the students to search for only “風味”. In this case, there are too many results, and it is difficult to find examples that are similar to the original Japanese sentence. Again, we discuss how we can adjust our search term to get closer to what we are looking for. Keeping the “の” (*no*) particle at the end should improve the results. I ask the students to search for “風味の” and to try to find an example that has any other kind of flavor before it and any other kind of food after it. About two thirds of the way down the first page of results, we encounter some useful results, seen in Figure 11. Both of these match the syntax of the problematic expression. The students quickly understand that we can use these results to change our expression to “black tea flavored sweets”.

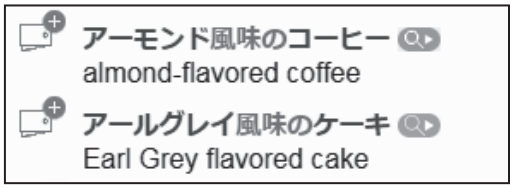


Figure 11. Helpful results from a search for *風味の* on *Eijiro on the WEB*

Through this example, the students see that if they get too few results, their search term might be too specific and they should think about how to shorten it. Likewise, if they get too many results, they should try to add to their search term to find more syntactically similar results. The “black tea flavored sweets” example is quite simple compared to other errors students might encounter. In order to use the site effectively, students need extensive practice with different types of syntactical problems.

To give them more open practice, I ask the students to write a paragraph on a given topic in Japanese. I choose different writing prompts depending on the level of the students. Once they have completed this task, I ask them to translate their writing into English on their own as best they can, or with the help of a dictionary. These translations generally contain many errors, and the students are roughly able to point out which parts were problematic for them. I help them to pinpoint particular errors, and then I have them attempt to use *Eijiro on the WEB* and *Weblio* to try to make corrections. This is usually quite difficult for students at first. It requires them to think about both their sentence construction and search methods. Many students are able to learn to use the sites effectively,

but some students struggle, and in some cases I have seen their writing become worse. I will discuss possible reasons for this in the next section.

Challenges

Overall, my experiences teaching *Eijiro on the WEB* and *Weblio* have been positive. Students are often impressed by the abundance of information they can find, and they quickly recognize the advantages that using these sites can have over their electronic dictionaries. However, students soon realize that using these sites effectively requires much more effort than simply typing a word into a dictionary. Some students learn how to use the sites very quickly, and most other students learn to use them with further practice and training. But in several cases that I have observed, some students' writing becomes worse when using these sites.

Larson-Hall (2015) looked specifically at teaching students to use *Eijiro on the WEB* to correct errors in their own writing and had similarly mixed results. She compared the error correction abilities of a focus group of 17 students trained to use *Eijiro on the WEB* with a control group of 17 students who did not receive training. While some students from the focus group were able to use *Eijiro on the WEB* to fix mistakes effectively, other students were not. Due to these varied results, the focus group did not perform statistically better than the control group. There are several reasons why I believe certain students struggle to use the sites.

The first reason, which I have already mentioned, is that using these sites is much more time consuming than simply looking up a word in a dictionary. Almost all of my students use an electronic dictionary in their classes, and this seems to be the norm in Japan. Because of their experience with using dictionaries, adapting to a new type of resource, such as a parallel corpus, is difficult. With a dictionary, students expect to type in a word and get a quick "answer." Finding the meaning of a word in a dictionary requires very little effort, and students have become accustomed to this. Similarly, quick machine translation sites such as Google Translate have become popular recently, and I have seen many of my students use them for writing composition. Even popular smartphone applications such as LINE have integrated quick machine translation functions. Students can simply type a word into a LINE Translate chat session, and receive an immediate "reply" of a single word translation. While dictionaries at least offer translations for several different meanings of a word, LINE translate

only gives one translation without any context or word form information. Unfortunately, many students seem to think this is adequate, and their resulting compositions are often incomprehensible. Students who fail to use *Eijiro on the WEB* effectively seem to treat the corpus simply as a dictionary, and this often leads to poor word choice. Since the corpus is built from authentic texts, it contains a lot of slang and colloquial language, so taking the time to look at usage is crucial. As shown earlier in Figure 4, the topmost results from a single word search in *Eijiro on the WEB* are just like those in a bilingual dictionary. Students who do not put time into analyzing word usage in sentences might randomly choose a translation from the topmost results. This can lead to some very strange word choices.

Another issue that leads to ineffective use of these sites is the continued prevalence of grammar-translation English instruction in Japan. This outdated style of instruction teaches students mechanical translation, and assumes a one-to-one relationship between the two languages. My own students are required to take a "Reading and Translation" course, which has them memorize lists of fixed sentence translations in English and Japanese. The students are then tested on their ability to reproduce these exact translations accurately. This limits awareness of the true one-to-many relationship between the two languages; as a result, many of my students have not yet developed the mind-set for the deeper syntactic analysis of language needed to use a parallel corpus effectively. However, while it may be challenging, exposure to a parallel corpus can be beneficial to students' linguistic awareness. Chujo et al. (2006) trained 72 Japanese students to use a Japanese-English parallel corpus to complete a variety of tasks such as looking for collocations and translating Japanese phrases into English. When asked an open ended question about what they learned by using the corpus, the students' most common answer was discovering that single words have many different translations.

Finally, students often have trouble pinpointing their own mistakes. They may feel a little unsure about full sentences or clauses, but they have difficulty knowing exactly which parts are problematic and need to be changed. My students have been much more able to correct mistakes that are indirectly pointed out. When checking papers, I often just underline or highlight the area of a mistake rather than making an explicit correction. Then I have the students use *Eijiro on the WEB* to attempt to identify and correct the mistake on their own. I have found that my students are usually able to fix their mistakes in this way. As mentioned earlier, my hope is that they will better retain the information because it requires active investigation. With this kind of practice, I expect them to get better at using the sites,

and in turn become more aware of their own mistakes.

Conclusion

Eijiro on the WEB and *Weblio* are powerful websites that can easily be applied in English language classes for Japanese students. Despite the initial difficulty for some students, training them to use these sites can lead to more natural writing, greater learner autonomy, and deeper linguistic awareness. More research into effective training is necessary, but these resources can be valuable tools for English language education in Japan.

References:

Boulton, A. (2009). Testing the limits of data-driven learning: language proficiency and training. *ReCALL, 21*(1), 37-54.

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language learning, 60*(3), 534-572.

Boulton, A. (2011). Data-driven learning: the perpetual enigma. In Goźdź-Roszkowski S (ed) *Explorations across Languages and Corpora*. Peter Lang, Frankfurt, pp.563-580.

Chujo, K., Utiyama, M., & Miura, S. (2006). Using a Japanese-English parallel corpus for teaching English vocabulary to beginning-level students. *English Corpus Studies, 13*, 153-172.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly* 34:548-560.

Davies, M. (2008-). The Corpus of Contemporary American English (COCA).

Ferris, D. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of second language writing, 8*(1), 1-11.

Hegelheimer, V. (2006). Helping ESL Writers Through a Multimodal, Corpus-based, Online Grammar Resource. *CALICO Journal*, 24(1), 5-32.

Johns, T., & King, P. (1991). Classroom Concordancing: English Language Research Journal, 4. *University of Birmingham: Centre for English Language Studies*.

Kosem, I. (2008). User-friendly corpus tools for language teaching and learning. In *Proceedings of the 8th teaching and language corpora conference* (pp. 183-192).

Larson-Hall, J. (2015). Using an online bilingual corpus dictionary to improve word choice and grammar. *Fukuoka JoGakuin Centre for the Study of English Language Teaching Journal* (英語教育研究センター), 3, 99-109.

Luo, Q. (2016). The effects of data-driven learning activities on EFL learners' writing development. *Springerplus,* 5(1), 1255-1268.

Luo, Q. & Liao, Y. (2015). Using Corpora for Error Correction in EFL Learners' Writing. *Journal of Language Teaching and Research*. 6(6). 1333-1342.

St. John, E. (2001). A case for using a parallel corpus and concordance for beginners of a foreign language. *Language Learning & Technology*, 5(3), 185–203.

Smart, J. (2014). The role of guided induction in paper-based data-driven learning. *ReCALL, 26*(2), 184-201.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning*, 46(2), 327-369.

Yoon, H. (2008). More than a linguistic reference: the influence of corpus technology on L2 academic writing. *Language Learning & Technology* , 12(2):31–48.

Yoon, H., & Hirvela A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, 13(4), 257-283.

Yoon H, Jo JW (2014) Direct and indirect access to corpora: an exploratory case study comparing students' error correction and learning strategy use in L2 writing. *Language Learning & Technology*, 18(1):96–117.

Websites mentioned in this article:

The Corpus of Contemporary American English (COCA) - https://corpus.byu.edu/coca/

Learner Corpora Association - http://www.learnercorpusassociation.org/

ALC's *Eijiro on the WEB* - https://www.alc.co.jp/

Weblio - https://ejje.weblio.jp/

Electronic Dictionary Project - http://www.eijiro.jp/

（Michael P. McGuire　英語国際学部講師）