Commentary on Four Studies for JALT Vocabulary SIG

Yukio Tono

*Advanced Research Unit, Graduate School of Global Studies,
Director of the World Language Center, Tokyo University of Foreign Studies*
doi: <https://doi.org/10.7820/vli.v08.1.Tono>

Abstract

The four papers in this volume focused on mainly three areas: readability (Pinchbeck), wordlist evaluation (Ishikawa; Culligan), and data-driven learning (McGuire). The author would argue that whilst applied corpus linguistics and L2 vocabulary research are closely related to each other, there has not been much interaction between the two disciplines. The review of the four papers has also shown such lack of interaction between the two fields. By reviewing each study, some additional concepts and previous related studies in corpus linguistics will be presented in order to fill those gaps.

Key words: corpus linguistics/ wordlist/ data driven learning/ parallel corpus

Introduction

Corpus linguistics used to be a paradigm supported by a small number of empirical linguists in Europe who were against Chomsky's mentalist approach (Sampson, 2005). Back in the 1960s, the sizes of corpora were tiny, around 1 million words at most, and only a large mainframe computer could process the data on magnetic tape using punch cards for inputting queries. Half a century has passed, and now the size of corpora available to general users can be twenty or thirty billion words, and one can handle a corpus of 100 million words in plain text files on a small laptop computer by using a user-friendly corpus query program. The Chomskyans' negligence in relying on empirical data has given way to more usage-based views of language, which revived the use of a large body of linguistic data as the mirror of language input and output that shapes the organization of an individual's or a language community's linguistic competence and usage (Bybee, 2006; Gries & Divjak, 2012; Sampson, 2009).

Now most researchers in theoretical and applied linguistics believe that corpora are a useful source of data that, if used correctly, shed light on various aspects of languages. L2 vocabulary acquisition/learning research is no exception. Due to the strong influence of Chomskyan linguistics, many second language acquisition (SLA) researchers worked primarily on syntax until the 1980s, when a gradual shift of attention from syntax to lexicon began. It was interesting that such pioneers in L2 vocabulary acquisition/learning research as Paul Meara, Paul

Nation, and Batia Laufer first paid little attention to corpora, partly because it was still too early then to realize its power. It was not until the growing influence of a balanced, representative corpus such as the British National Corpus (BNC) at the beginning of the 21st century that many applied linguists started to appreciate the power of mega-corpora. Another reason might be that many L2 vocabulary researchers were interested in vocabulary teaching and learning techniques or strategies, which seemed to be somewhat removed from what corpora can tell us about languages (Schmitt & McCarthy, 1997). It was Nobert Schmitt that started to use corpora intensively for L2 vocabulary research (Schmitt, 2010). His experience with Ronald Carter and Michael McCarthy in the Nottingham school helped him a great deal to appreciate the power of corpora for investigating L2 vocabulary research.

Among the corpus linguistics community, applications of corpora in language learning/teaching research started in the beginning of the 1990s, when the participants of the ICAME Conference, which was one of the first conferences for those who developed the Brown Corpus and its family, discussed the possibility of holding a conference dedicated to teaching and language corpora (TALC). The first TALC conference was held at Lancaster University in 1994 (Wichmann et al., 1997). I attended the second TALC in 1996, where I met numerous fantastic scholars at Lancaster University, such as Geoffrey Leech, Tony McEnery, Nick Smith, and Paul Rayson. It was around that time that the BNC was released in 1994 and made available within European countries only. In 1995, as fruit of the first 100 million balanced corpora, two principal monolingual learner's dictionaries, the *Longman Dictionary of Contemporary English* (LDOCE, the third edition) and the *Oxford Advanced Learner's Dictionary* (OALD, the fifth edition), were revised based on the analyses of the BNC. Together with COBUILD, the impact of corpora was such that it changed the views of language researchers and teachers toward the importance of 'big data' in language pedagogy. That was also about the time when I started to compile a corpus of learner English. The founder of the International Corpus of Learner English (ICLE), Sylviane Granger at Catholic University of Louvain, also launched the project ICLE at the beginning of the 1990s (Granger, 1998). Learner corpus research has been multiplying since then and has contributed to our understanding of lexico-grammatical features of learner language. This is a field of studies closely linked to L2 vocabulary research, but L2 vocabulary and learner corpus researchers did not talk to each other very often until recently.

The same kind of lack of interaction between the two fields can be seen in the present collections of papers as well. I notice that in corpus linguistics, some of the questions addressed by L2 vocabulary researchers have been investigated in more detail and sometimes provide more concrete ideas and heuristics concerning methodologies. This kind of gap can be and needs to be filled. Therefore, it would indeed be an excellent opportunity for the two fields to meet and discuss various issues at this JALT Vocabulary SIG.

The four papers in this volume focused on mainly three areas: readability (Pinchbeck), wordlist evaluation (Ishikawa and Culligan), and data-driven learning (DDL) (McGuire). In the following sections, I will discuss the four studies regarding the issues of wordlist evaluation, readability, and DDL, respectively.

The four studies

A Reconsideration of the Construct of “A Vocabulary for Japanese Learners of English”: A Critical Comparison of the JACET Word-lists and New General Service Lists by Shin’ichiro Ishikawa

Researchers have taken different approaches toward wordlist evaluation. There are two main types. One is an internal validation approach, which is to evaluate the wordlist internally by examining the validity of sampling and representativeness of the source data based on which the wordlist was created. Ishikawa (in this volume), for instance, examined two New General Service Lists (NGSLs) (Brezina & Gablasova, 2015; Browne, Culligan, & Phillips, 2013) and their source data in his review. The sources for the two wordlists happened to come from native speaker (NS) corpora only (see Table 2 in Ishikawa), and Ishikawa argued that there might be a question of whether a wordlist based solely on NSs’ usage is valid as a pedagogical list for language learners.

The other approach is an external validation approach, in which wordlists are compared against each other regarding the coverage of a particular text. For this, one could evaluate the quality of the wordlist by comparing the coverage of texts of different genres or text types against the wordlist (Chujo 2012; Chujo & Oghigian, 2009). One could also cross-compare different wordlists. A good example is Nation (2004), in which a comparison between the General Service List (GSL) and the BNC frequency list revealed the characteristics of each list (e.g. educational and dated for the GSL and the British nature of the BNC) due to their different text sources. Ishikawa used this method and compared the new JACET 8000 against two NGSLs. Culligan (in this volume) is unique in this sense because he uses new techniques (i.e., responses to Yes/No tests scaled using Item Response Theory (IRT)) to assess the contents of the wordlists.

Ishikawa described the history of the development of JACET 8000, arguing for the value of blending frequency information from corpora and intuition-based lists. Also, he questioned the validity of wordlists based only on NS corpora. In his case study, he compared JACET 8000 against two NGSLs and found that while there is much core vocabulary covered in the three lists, NGSLs tend to include words denoting “business and higher-order mental activities,” which are missing in JACET 8000. On the other hand, JACET 8000 contains words related to personal and daily life situations and morphologically much shorter words.

The results seem to show a difference in what source texts were used for wordlist creation: NS corpora only versus other sources. The two NGSLs mainly use NS corpora (cf. Table 2 in Ishikawa) while JACET 8000 was based on various materials relevant to English learners in Japan. The fundamental concept of corpus compilation has to be revisited here, where such concepts as the target population, balance, representativeness, and sampling are to be discussed carefully in order to fulfill the aims of target wordlist construction. The NGSLs aim to serve English learners in general, which is the main reason why NS corpora were used for word selection. However, it could be argued that the wordlist merely reflects the characteristics of vocabulary regarding words used most frequently by NSs, which may not be the same as the learning order of L2 vocabulary. Ishikawa argues that an educational wordlist should have a clear target user and use in mind,

to which I could not agree more. The GSL sounds good, but what is meant by “general service” has to be critically evaluated. In this sense, I hasten to add, Ishikawa’s argumentation about the definition of “vocabulary for Japanese learners of English” may need further scrutiny, regarding proficiency levels, situations of use, and target learners’ age among other things.

Related to Ishikawa’s paper, it should be noted that recently more dynamic models of wordlists have been proposed. For example, the English Vocabulary Profile (EVP; Capel, 2015), created based on analysis of the Cambridge Learner Corpus, shows the difficulty level information based on the Common European Framework of Reference for Languages (CEFR; A1 to C2) not only for each headword but also for each word sense, colloquial expressions, and phrasal verbs. While how to determine the CEFR levels of word meanings is still an empirical question, this wordlist sheds new light on the direction of pedagogical wordlist creation.

Another wordlist is the CEFR-J Wordlist (Tono, 2013). It contains 7500 words covering CEFR levels A1 to B2. This wordlist was based on the analysis of English textbooks used in China, Korea, and Taiwan, together with the information from the CEFR-related resources. The motivation behind this project is to compile a wordlist reflecting current best practice in Asian countries/regions, where English has been taught at elementary school, as well as some pedagogical implications from the CEFR-based lexical resources such as the EVP, Threshold Level (van Ek & Trim, 1990), and the Core Inventory for General English (North, Ortega, & Sheehan, 2010). The CEFR-J Wordlist provides the information about nouns for their general versus specific notions indicated in Threshold Level as well as thematic categories from the Core Inventory for General English. This information is vital as a word is used in conjunction with certain constructions. Together they serve to perform a specific function in given contexts or situations. This specification of language points is realized, in the CEFR terms, by a set of illustrative functional descriptors that will produce certain situations and tasks where grammar and vocabulary are united to produce utterances. Vocabulary should be seen as a realization of functions in context. Thus, this kind of wordlist does not stand on its own. It should work most effectively as part of a broader framework and other linguistic inventories such as a grammar profile or text profile (Tono, 2017).

Evaluating Corpora with Word Lists and Word Difficulty by Brent A. Culligan

Culligan proposed a unique method of evaluating the quality of different wordlists regarding the average IRT difficulty estimates for the words on the wordlist. The title of the paper contains the phrase “evaluating corpora,” but actually what he did was another type of external validation of the wordlists. Although I am not entirely convinced whether a sufficient number of data points were obtained for each lemma from Yes/No tests in order to produce reliable theta scores, the results still show some interesting observations.

Firstly, the results show that there is considerable overlap in word difficulty between the GSL 2000 and the Academic Word List (AWL), which suggests that the difference between the GSL 2000 and the AWL was not as clear as was intended. It is possible that many words added in the AWL could have been in the

GSL 2000 if more texts had been added to the original data examined by Michael West. The purpose of the AWL is to supplement the GSL with general-purpose academic words found in various academic texts, which assumes that a group of words in the AWL should not be too domain-specific. In this sense, this similarity between the GSL 2000 and the AWL might indicate that the control of word difficulty levels for the GSL 2000 (-0.79 logits, $SD = 2.03$) may not be so successful, compared to that of the NGSL (-1.59 , $SD = 1.68$).

Secondly, Culligan concludes that “this analysis provides evidence that seems to indicate researchers that use a word family approach, particularly using word affixation beyond Level 2 will result in more difficult word lists.” I could not agree with this claim more, because in English corpus linguistics at least, we have never used word family as a solid counting unit for a word. Also, as one scrutinizes the GSL entries, there are many cases where lemmas and word families were not treated consistently. Some of the decisions made by Michael West seem to be arbitrary and need fine-grained reanalysis. In this sense, the approach adopted by the NGSL team, both in the UK and Japan, seems to be reasonable, since it is not a good idea to follow the GSL blindly.

Methodologically, Culligan’s study suggests an interesting direction of evaluating the wordlists in light of learners’ perception of word familiarity. While what is known in Yes/No tests is still very partial, it would certainly provide useful information about how the wordlists behave against such measures. This reminds me of the recent development of Pearson’s Global Scale of English (de Jong et al., 2016) and its accompanying word and grammar resources. Their attempt at estimating word difficulty levels on the GSE 10 to 90 scales is also an excellent example of applying logit scores converted to the GSE scores. In their survey, however, the word difficulty level was assessed by the teacher questionnaire survey, in which teachers were asked whether a particular word is necessary to learn at a given CEFR level. It would be interesting to compare the results of Culligan’s study to the word difficulty level proposed by the GSE.

Another interesting issue as a follow-up study would be to investigate whether the results of Yes/No tests will be affected by Japanese learners’ prior knowledge of katakana loan words. Many entries in the GSE and other wordlists are already part of Japanese katakana vocabulary, and once students can recognize the correspondence between English spellings and their katakana equivalents, they can tell the meaning in many cases. Thus, the item difficulty scores might be affected by such knowledge as recognizing katakana words as cognates in the test items. Also, this judgment of cognates is heavily affected by L2 learners’ meta-cognitive knowledge, and thus, the target population of Japanese learners of English might yield a different result.

Validating the Construct of Readability in EFL Contexts: A Proposal for Criteria by Geoffrey G. Pinchbeck

Pinchbeck’s paper focused on the lexical component of readability and proposed a list of criteria to evaluate candidate components of readability formulas for Japanese learners, including (a) transparency in calculating a readability index, (b) availability, (c) non-commercial, (d) accessible technology, (e) diagnostic, and (f) empirically valid measures. They all look reasonable, and some of the

readability measures available in Japan (e.g., Yomiyasusa Level and Kyoto Scale) need further refinement in these respects.

Pinchbeck relies on cloze tests as the basis of readability scales, but I am not entirely convinced whether what is measured by cloze tests is equivalent to text difficulty. It is not clear that one can define readability by the results of cloze tests. Cloze tests measure various aspects of linguistic competence: grammatical, lexical, collocational, and discursal knowledge. While these different types of knowledge are closely related to learners' linguistic competence, learners do not usually require them as they read. Most test items asked in cloze tests tap on L2 learner's productive knowledge. Readability, however, is more closely linked to the difficulty level of receptive knowledge. I would argue that assessing readability levels using cloze tests provides only an indirect link to text difficulty as a construct and that a more valid and clear operational definition of readability is needed.

Most readability formulas developed so far are based on simple surface measures such as the total text length, the average word or sentence length, and the word difficulty level. In corpus linguistics, we can find other useful substitutes for existing readability measures. For example, Tom Cobb's Compleat Lexical Tutor (Cobb, 2018) provides various measures for text characteristics based on different wordlists such as the GSL, the BNC, or the Corpus of Contemporary American English (COCA). Many concordancers such as WordSmith (Scott, 2018), AntConc (Anthony, 2018b), and CasualConc (Imao, 2018) have functions for measuring text statistics. Text Inspector developed by the English Profile team can produce various readability measures and their original CEFR level estimation.

Our project called the CEFR-J (Tono, 2017) aims to construct a CEFR-based original framework together with various profile information including the CEFR-J Wordlist, Grammar Profile, and Text Profile. There is an online CEFR level assignment tool called the CEFR-based Vocabulary Level Analyzer, developed by Satoru Uchida (see Figure 1). When you select a text whose CEFR level you want to know, the tool will produce an estimated CEFR level based on the regression model of different CEFR-J word proportions, existing readability measures, the average number of verbs per sentence, and the ratio of B-level words to A-level words.

These new metrics are all produced as a part of reference level descriptions for the CEFR. This will inform L2 vocabulary researchers of new types of lexical and grammatical profile.

Toward Written Error Correction with a Japanese–English Parallel Corpus: DDL in the Japanese EFL Classroom by Michael McGuire

Data-driven learning is becoming popular as an inductive approach to learning a language. McGuire rightly commented that there are very few studies on actual DDL activities, primarily related to written error correction. He went on to suggest that the use of Japanese–English parallel corpora would make a DDL approach more accessible and useful. Regarding the effects of DDL on error correction in L2 writing, Tono, Satake and Miura (2013), and Satake (2018) tackle the very issues of how DDL helps L2 learners identify and correct errors. Satake (2018) found that prior identification of errors by teachers would facilitate

CVLA: CEFR-based Vocabulary Level Analyzer (ver. 1.1)

[Legend]

A1:example, A2:example, B1:example, B2:example, C1:example, C2:example,

NA content words:example, NA others:example

#Small numbers indicate the ranking in COCA (added only to nouns, verbs, adjectives, and adverbs).

[Input]

Using₉₉ computers₆₁₃, it is not₂₂₇₂₇ currently₁₇₅₇ possible₄₇₃ for any practitioners₄₀₈₃ to easily₁₃₀₄ and accurately₅₁₁₈ analyse₁₁₉₀₉ the deeper structural₃₈₃₁ and semantic₉₉₄₈ features₁₂₂₁ of texts₁₃₁₆ (i.e., dialect₁₁₀₄₆; rhetorical₇₅₉₇ structure₉₅₀; topic₁₉₉₇ background₁₄₉₉, specificity₁₁₁₀₀ and complexity₃₈₁₅; cohesive₁₃₄₁₄ devices₁₅₃₉; allusion₁₂₁₀₄; metaphor₄₀₅₈; the level₃₁₆ of inferencing required₄₃₉ of the reader₁₂₇₅; and/or formulaic₂₂₆₂₅ language₆₇₄; etc.); for this we must await₃₉₃₀ further advances₂₄₄₁ in natural₇₂₁ language₆₇₄ processing₃₂₆₉ technologies₅₆₃. In the meantime₄₈₅₇, however₂₇₄, I propose₁₇₈₃ a re-examination₃₂₃₁₄ of the lexical₁₈₀₁₀ component₁₇₂₅ of readability₂₂₄₆₇ is warranted₆₂₈₅, particularly₈₀₈ in Japan and in other₇₇ EFL teaching₁₁₅₆ contexts₁₃₈₃. It is my hope₁₂₃₇ that this editorial₅₇₄₂ might generate₁₈₀₈ some interest₄₃₀ towards the development₄₂₇ of large-scale₆₂₁₃ validity₄₂₂₀ studies₂₁₄ that would develop₅₂₀ new₉₁ pedagogical₁₀₇₄₀ tools₁₂₇₆ for more valid₄₈₁₀ and efficient₃₁₃₉ matching of Japanese₄₁₃₉ K-12 and/or adult₇₆₈₂ EFL learners₄₆₂₀ to English₁₈₄₇ texts₁₃₁₆.

[125 words, mean rank: 4879.62, min: 77, max: 32314]

Lexical Index	Raw Frequency	Explanation
---------------	---------------	-------------

Mode: R				
CEFR	ARI	VperSent	AvrDiff	BperA
A1	5.73	1.49	1.31	0.08
A2	7.03	1.82	1.41	0.12
B1	10.00	2.37	1.57	0.18
B2	12.33	2.88	1.71	0.26
Input	26.65	3.67	2.25	0.60
Estimated level	C2	C2	C2	C2

Estimated Text Level:C2

Figure 1. The CEFR-based Vocabulary Level Analyzer (Uchida, 2018).

learners' correction of those errors using corpora. However, there is an interaction between the ease of error corrections and the types of errors; omission and addition errors were more easily corrected than misinformation errors, due to the saliency in error status on concordance lines.

McGuire also encouraged the use of parallel corpora and Eijiro on the Web as a kind of parallel corpus. This issue is closely related to the three challenges for using DDL in the classroom that Laurence Anthony mentioned in his plenary speech given at Asia Pacific Corpus Linguistics Conference (APCLC) 2018 (Anthony, 2018a). The first challenge is to find a suitable corpus. In order to make a useful DDL material, it is essential to adjust the difficulty level of the texts on concordance lines to the level of target learners. McGuire recommends the use of parallel corpora to bridge the gap of learners' lack of understanding of difficult examples. A similar approach has been already taken in previous projects such as Chujo and her colleagues (Chujo et al., 2006).

DDL researchers usually use various corpus building tools to compile so-called DIY corpora. For example, AntFileConverter (Anthony, 2017) can convert PDF or Word files to plain texts, so if supplementary teaching materials are available in PDF or Word format, it can add those texts to the DIY corpus for specific teaching purposes. People also use tools such as BootCat, Sketch Engine, and WordSmith to crawl the Internet and collect web materials automatically.

Secondly, Anthony (2018a) brought up the issue of interacting with a corpus, which involves how to interpret patterns of use found on KWIC lines of a particular search word, in the information about word frequency and distribution, and in a list of collocations and related association measures. This issue is complicated, because the purposes of DDL are diverse, as it could aim for convergent learning, in which learners are expected to come to the same generalization through the interpretation of corpus evidence, while it could also aim for more divergent learning, where each learner will make their own discoveries through data exploration in a serendipitous manner (Leech, 1997).

The final challenge mentioned by Anthony (2018a) was using corpus tools. In order to have an active DDL environment, it would be necessary for learners in the classroom to have hands-on experience with corpus data. Printed versions of corpus data have some positive effects, but many corpus query tools need a certain degree of mastery for flexible data extraction and learners need to overcome computer literacy problems if they are really serious about exploring the data in a corpus. The development of user-friendly corpus tools such as SKELL (Baisa, 2014) and SCoRE (Chujo, Oghigian, & Akasegawa, 2015) is quite useful in this sense.

McGuire recommends the Eijiro on the Web because it has a function for showing lists of example sentences with Japanese translations in Key Word In Context (KWIC) format. Many electronic dictionary interfaces such as Eijiro or CASIO's pocket dictionaries now have this function, which looks similar to corpus data. However, it should be noted that the number of examples in the Eijiro or other electronic dictionaries may not be large enough to show significant patterns of use of a particular search word in context. Thus, learners need to be warned about how to interpret the example list on such systems.

Overall, the effects of DDL have to be evaluated in light of the role it plays in the lesson and to answer such questions as whether DDL is used for presenting a new structure or vocabulary, practicing vocabulary or grammar, or for more task-based production activities. Many DDL studies did not seem to take into account the overall syllabus contents and how DDL works within that syllabus. The inductive approach sounds fine, but it may take too much time to use DDL for every stage of a lesson. Then, one should seriously ask what aspect of learning (e.g., comprehension vs. production; lexis and grammar; and general vs. ESP) must be most effectively delivered using DDL.

Some Suggestions for Future Research: The Interface between Grammar and Lexis

Concerning McGuire's suggestion to use parallel corpora for DDL, I will finish my commentaries by making a brief report on the application of parallel learner corpora for finding Japanese EFL learners' overuse and underuse of grammatical structures. In L2 vocabulary acquisition research, there are many studies on wordlists, but not many studies have been done to summarize the overall usage patterns of grammar constructions. Since there is no longer a clear boundary between lexis and grammar, more research needs to be conducted to profile L2 learners' use of grammar items.

The Japanese EFL Learner Corpus (JEFLC Corpus; Tono, 2006) contains English essays written by 10,038 junior and senior high school students in Japan (the total number of running words: 669,304 tokens). All the essays were proofread by NSs, thus making the original version comparable against its corrected version. A set of pattern-matching queries for extracting grammar items was prepared and automatically extracted frequencies of grammar items from both the original and the corrected versions of the JEFLC Corpus (see Table 1).

By extracting all the frequencies of each grammar item from the original and corrected versions of the JEFLC Corpus, we can contrast the two versions regarding the overuse versus underuse of each grammar item, shown in Figure 2. The JEFLC Corpus is classified by the CEFR levels, which makes it possible to observe the transition of overuse and underuse phenomena of a particular grammar item throughout the CEFR levels.

Table 1. Extraction rules for CEFR-J grammar profile¹

ID	Item	Pattern
26	INDEFINITE PRONOUN: none	\none_NN_none\b
49	COMPARATIVE and COMPARATIVE (the same adjective)	\b(\S+_(JJR RBR)_\S+) and_CC_and \1
66	TENSE/ASPECT: PAST PROGRESSIVE (AFFIRMATIVE DECLARATIVE)	(was were)_VBD_be(?!(going_VVG_go_to_TO_to gonna_VVG_gonna) \S+_V._\S+) \S+_V.G._\S+
145	AUX+PERFECT (AFFIRMATIVE DECLARATIVE)	(?!cannot\b)\S+_MD_\S+have_VH_have \S+_V.N._\S+

Shorthand Code	RAW FREQ.										RELATIVE FREQ. (per mil. words)				
	A1_origin	A1_correc	A2_origin	A2_correc	B1_origin	B1_correc	B2_origin	B2_correc	A1_origin		A1_correc	A2_origin	A2_correc	B1_origin	B1_correc
PP.I.am	639	550	944	1,009	533	583	25	31	■	■	4,858	3,574	3,049	2,979	2,512
PP.I.am	26	41	68	81	38	49	2	3	■	■	198	266	220	239	179
PP.you_are	5	8	25	26	28	35	2	2	■	■	38	52	81	77	132
PP.are_you	9	9	5	10	10	12	3	3	■	■	68	58	16	30	47
PP.he.she_is	162	56	277	184	159	105	7	5	■	■	1,232	364	895	543	749
PP.we_are	65	24	132	72	96	71	0	0	■	■	494	156	426	213	452
PP.they_are	86	143	250	365	101	157	7	6	■	■	654	929	808	1,078	476
PP.they_are_not	1	3	19	24	8	11	0	0	■	■	8	19	61	71	38
PGEN	2,758	3,735	6,748	7,715	4,763	5,022	177	189	■	■	20,969	24,271	21,799	22,779	22,450
PPO	770	1,183	2,795	3,480	2,286	2,610	97	109	■	■	5,854	7,687	9,029	10,275	10,775
DT.this.that_is	85	100	208	345	168	257	3	5	■	■	646	650	672	1,019	792
DT.these.those_are	9	2	30	15	17	7	1	1	■	■	68	13	97	44	80
PP.it_is	654	408	1,287	1,080	649	557	26	23	■	■	4,972	2,651	4,158	3,189	3,059
PP.it_is_not	23	18	90	74	60	49	3	1	■	■	175	117	291	218	283
DT.this.that_N	294	390	776	818	599	626	39	34	■	■	2,235	2,534	2,507	2,415	2,823
DT.these.those_N	11	45	56	110	45	75	2	3	■	■	84	292	181	325	212
DT.a.an	2,078	4,461	5,398	8,721	3,963	5,855	164	228	■	■	15,799	28,989	17,438	25,749	18,679
DT.the	2,386	5,143	7,515	11,228	6,049	8,216	228	279	■	■	18,141	33,421	24,276	33,151	28,512
DT.some.any	412	472	1,069	1,141	670	698	52	53	■	■	3,132	3,067	3,453	3,369	3,158

Figure 2. The overuse versus underuse of grammar items in the JEFLL parallel corpus

Table 2. The 10 most underused grammar items

Lexical forms	A1	A2	B1	B2	Average
<i>being</i> + PAST PARTICIPLE	0.000	0.256	0.334	0.000	0.147
PREP+RELATIVE PRONOUN	0.090	0.283	0.226	0.000	0.150
TENSE/ASPECT: PRESENT PERFECT	0.268	0.398	0.415	0.000	0.270
RELATIVE PRONOUN: NONRESTRICTIVE	0.047	0.322	0.303	0.525	0.299
MODAL/AUX: <i>should</i>	0.000	0.547	0.691	0.000	0.310
RECIPROCAL PRONOUN: <i>each other</i>	0.731	0.421	0.115	0.000	0.317
TENSE/ASPECT: PRESENT PERFECT PROGRESSIVE	0.146	0.505	0.668	0.000	0.330
<i>whether</i>	0.351	0.469	0.534	0.000	0.338
WH-QUESTION: <i>When...?</i>	0.585	0.377	0.420	0.000	0.345
PASSIVE: PAST PERFECT	0.123	0.729	0.551	0.000	0.351

Table 3. The 10 most overused grammar items

Lexical forms	A1	A2	B1	B2	Average
<i>These/Those are...</i>	5.265	2.188	2.594	1.050	2.774
<i>He/she is...</i>	3.385	1.647	1.617	1.470	2.030
<i>There + be...</i>	1.404	1.810	1.978	2.100	1.823
<i>It is not...</i>	1.495	1.331	1.308	3.150	1.821
ELLIPTICAL ACCUSATIVE RELATIVE PRONOUN	1.209	1.245	1.434	2.071	1.489
<i>Much</i> + UNCOUNTABLE NOUN	1.083	1.189	1.576	2.100	1.487
MODAL/AUX: <i>would</i>	0.439	0.574	0.518	4.201	1.433
COMPARATIVE OF SUPERIORITY: <i>more</i> + ADJ/ADV	0.794	0.802	0.867	3.150	1.403
<i>It is...</i>	1.875	1.304	1.244	1.187	1.403
TENSE/ASPECT: PRESENT (BE)	1.791	1.320	1.248	1.214	1.393
TENSE/ASPECT: PRESENT (BE): NEGATIVE	1.390	1.346	1.258	1.470	1.366

Tables 2 and 3 show the 10 most underused and overused grammar items, respectively. The number in each cell shows the ratio of the number of grammar items in the original essay to that in the corrected essay. If the number is below 1, it means the item is underused. If it is above 1, it is overused. Our analysis shows that Japanese junior and senior high school students tend to underuse grammar items rather than overuse them. Underused items are typically those constructions which involve more than one grammar item, for example, ‘being + past participle [PROGRESSIVE+PASSIVE]’.

The detailed corpus analysis of such overuse/underuse grammar patterns at each CEFR level will help materials developers focus on those grammar items that need special attention at different CEFR levels and change the organization of grammar textbooks drastically. Many of these grammar items involve the use of essential verbs and function words, which will shed new light on the treatment of core vocabulary more systematically.

Conclusions

This review convinces me even more that there is a very close relationship between applied corpus linguistics and L2 vocabulary research. The four studies reviewed in this volume clearly show some directions toward more communication between the two fields. As was mentioned in the introduction, corpus linguists working on learner language or pedagogical materials sometimes do not realize the background of SLA, which prevents them from interpreting the corpus analysis results in a way that is meaningful to SLA researchers. Likewise, SLA researchers are not fully aware of the wealth of resources and tools available in corpus linguistics and natural language processing, which will greatly facilitate data analysis. This has been said many times, but I have the impression that the more corpus and Natural Language Processing (NLP) technologies advance, the wider the gap will be for those who try to fill it. I would recommend that those colleagues who are active in an interdisciplinary Special Interest Group (SIG) like JALT Vocabulary (JALT VOC) should promote the exchange of ideas and people even more to help bridge the gap.

Note

1. This grammar pattern extraction was mainly done by Yasutake Ishii. See Ishii and Tono (2018) for more detail.

References

- Anthony, L. (2017). *AntFileConverter* (Version 1.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from <http://www.laurenceanthony.net/software>
- Anthony, L. (2018a). Understanding and advancing the Data Driven Learning (DDL) approach. In Tono, Y. and Isahara, H. (eds.) *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference* (pp. 1–4). Asia Pacific Corpus Linguistics Association. Retrieved from <https://apclc2018.org/proceedings/>
- Anthony, L. (2018b). *AntConc* (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from <http://www.laurenceanthony.net/software>
- Baisa, Vít a Vít Suchomel. (2014). SkELL – Web Interface for English Language Learning. In *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing* (pp. 63–70). Brno: Tribun EU.
- Brezina, V. & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics* 36(1), 1–22.
- Browne, C., Culligan, B., & Phillips, J. (2013). *New academic word list 1.0*. Retrieved from <http://www.newgeneralservicelist.org/>
- Bybee, J. (2006). *Frequency of use and the organization of language*. New York: Oxford University Press.

- Capel, A. (2015). The English Vocabulary profile. In J. Harrison & F. Barker (eds.), *English profile in practice* (pp. 9–27). Cambridge: Cambridge University Press.
- Chujo, K. (2012). Corpus no Eigo Kyoiku heno Jissen Riyou ni mukete. *Proceedings of the 42nd English Language Education Society of Japan*, (pp. 1–8).
- Chujo, K. & Oghigian, K. (2009). How many words do you need to know to understand TOEIC, TOEFL & EIKEN? An examination of text coverage and high frequency vocabulary. *The Journal of Asia TEFL*, 6(2), 121–148.
- Chujo, K., Oghigian, K. & Akasegawa, S. (2015). A corpus and grammatical browsing system for remedial EFL learners. In A. Leńko-Szymańska & A. Boulton (eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 109–128). Amsterdam: John Benjamins.
- 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.
- Cobb, T. (2018). *Compleat Lexical Tutor*. (Version 8.3) [Computer software] Retrieved from <https://www.lex tutor.ca/>
- de Jong, J. H. A. L., Mayor, M. & Hayes, C. (2016). *Developing Global Scale of English Learning Objectives aligned to the Common European Framework*. Retrieved from <https://online.flippingbook.com/view/894261/>
- Granger, S. (ed.) (1998). *Learner English on computer*. Harlow: Addison Wesley Longman.
- Gries, S. T. & Divjak, D. (eds.) (2012). *Frequency effects in language learning and processing* (Trends in Linguistics. Studies and Monographs). Berlin: Mouton de Gruyter.
- Imao, Y. (2018). *CasualConc*. (Version 2.1.1) [Computer software] Retrieved from <https://sites.google.com/site/casualconcj/>
- Ishii, Y. & Tono, Y. (2018). Investigating Japanese EFL learners' overuse/underuse of English grammar categories and their relevance to CEFR levels. In Y. Tono & H. Isahara (eds.), *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference* (pp. 160–165). Asia Pacific Corpus Linguistics Association. Retrieved from <https://apclc2018.org/proceedings/>
- Leech, G. (1997). Teaching and language corpora: A convergence. In Wichmann, et al. (eds.) (1997), 1–23.
- Nation, P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 3–13). Amsterdam: John Benjamins.
- North, B., Ortega, A. & Sheehan, S. (2010). *A core inventory for general English*. British Council and EAQUALS. Retrieved from <https://englishagenda.britishcouncil.org/sites/default/files/attachments/pub-british-council-eaquals-core-inventoryv2.pdf>
- Sampson, G. (2005). *The "Language Instinct" Debate*. London: Continuum.

- Sampson, G. (2009). A linguistic axiom challenged. In G. Sampson, D. Gil & P. Trudgill (eds.), *Language complexity as an evolving variable* (pp. 1–18). Oxford: Oxford University Press.
- Satake, Y. (2018). *The effects of corpus use on error correction and error identification in L2 writing*. Unpublished Ph.D. dissertation. Tokyo: Tokyo University of Foreign Studies.
- Schmitt, N. (2010). *Researching vocabulary*. London: Palgrave Macmillan.
- Schmitt, N. & McCarthy, M. (eds.), (1997). *Vocabulary: Description, acquisition and pedagogy*. Cambridge: Cambridge University Press.
- Scott, M. (2018). *WordSmith Tools*. (Version 7) [Computer software] Retrieved from https://lexically.net/downloads/version7/wordsmith_versions.htm
- Tono, Y. (ed.) (2006). *Nihonjin Chukousei I-man nin no Eigo Corpus: The JEFLL Corpus*. (The JEFLL Corpus Handbook) Tokyo: Shogakukan.
- Tono, Y. (ed.) (2013). *The CEFR-J Handbook*. Tokyo: Taishukan.
- Tono, Y. (2017). The CEFR-J and its impact on English Language Teaching in Japan. *JACET International Convention Selected Papers*, 4, 31–52. Tokyo: Japan Association for College English Teachers.
- Tono, Y., Satake, Y. & Miura, A. (2013). The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL*, 26(2), 147–162. doi:10.1017/S095834401400007X.
- Uchida, S. (2018). *CVLA: CEFR-level Vocabulary Level Analyzer* (Version 1.1) [Computer software] Retrieved from <http://dd.kyushu-u.ac.jp/~uchida/cvla.html>
- van Ek, J. A. & Trim, J. (1990). *Threshold Level 1990*. Council of Europe Publishing.
- Wichmann, A., Fligelstone, S., McEnery, T. & Knowles, G. (eds.) (1997). *Teaching and Language Corpora*. Harlow: Addison Wesley Longman.