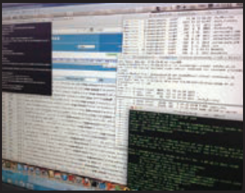
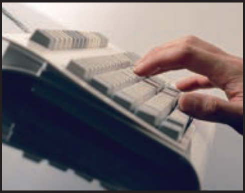
JAECS
Japan Association for English Corpus Studies

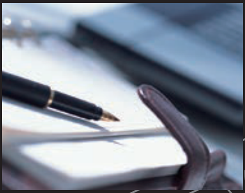
2025

英語コーパス学会

ISSN 1340-301 X





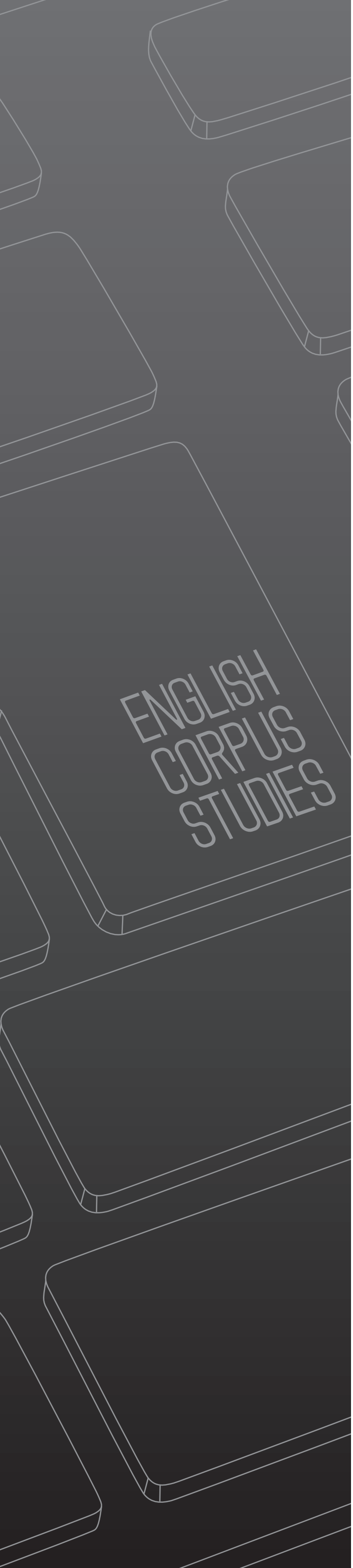


英語コーパス研究

第32号

ENGLISH
CORPUS
STUDIES

English Corpus
Studies 32



2025年度
英語コーパス学会役員

会　　長：田畑　智司

副 会 長：石井　康毅（研究促進担当）
　　　　　水本　　篤（学会誌担当）
　　　　　小島ますみ（総務担当）
事務局長：小島ますみ
事務局員：八野　幸子，石井　達也（会計担当）
　　　　　木山　直毅，森下　裕三（広報担当）
　　　　　佐竹　由帆（総務担当）
理　　事：石井　康毅，今尾　康裕，今林　　修，内田　　諭，
　　　　　岡田　　毅，金澤　俊吾，加野まきみ，小島ますみ，
　　　　　小林雄一郎，佐竹　由帆，杉森　直樹，住吉　　誠，
　　　　　田畑　智司，塚本　　聡，投野由紀夫，水本　　篤，
　　　　　山﨑　　聡，山﨑のぞみ（18名）
幹　　事：阿部真理子，五百藏高浩，石井　達也，石川　有香，
　　　　　和泉　絵美，宇佐美裕子，鎌倉　義士，木山　直毅，
　　　　　長　加奈子，西垣知佳子，仁科　恭徳，八野　幸子，
　　　　　藤原　康弘，森下　裕三，渡辺　拓人（15名）
監　　査：伊藤　亮太
顧　　問：赤野　一郎，中村　純作，堀　正広

■　学会誌編集委員会
＜委 員 長＞水本　　篤
＜委　　員＞池尾　玲子，内田　　諭，鎌倉　義士，菊池　翔太，
　　　　　　木山　直毅，福元　広二，水本　　篤，山﨑のぞみ

■　学会賞選考委員会
＜委 員 長＞今林　　修
＜委　　員＞今尾　康裕，今林　　修，杉森　直樹，谷　　明信，
　　　　　　長谷部陽一郎，山﨑のぞみ

■　大会実行委員会
＜委 員 長＞田畑　智司
＜委　　員＞田畑　智司，今尾　康裕，小島ますみ，菅原　裕輝，
　　　　　　鈴木　大介，南澤　佑樹，森下　裕三

ISSN 1340-301 X

英語コーパス研究

第 32 号

英語コーパス学会

2025

目　　次

論文

研究ノート

ソフトウェア紹介

「論文」

Cues to identifying support verb constructions: A corpus-based study of [Verb + EXAMINATION]

Taishi CHIKA and Kazuho KAMBARA

Abstract

This paper describes a corpus survey of [Verb + EXAMINATION] to explore the formal specifications of support verb constructions (SVCs). Previous studies described the properties of SVCs by focusing on the apparent semantic and syntactic paucity of support verbs (Brugman, 2001; Newman, 1996; Wierzbicka, 1982) and the process of argument transfer triggered by the complements (Grimshaw & Mester, 1998; Grimshaw, 1990). However, these verb-centered approaches face the issue of ambiguity between light and heavy senses in context (e.g., *make$_{light}$ a diagnosis*, *make$_{heavy}$ a certificate*), and coverage of low frequency support verbs (e.g., *sustain an injury*). To address these issues, we point out the need for formal specifications of SVCs targeted on their compliments. Our corpus analysis of [Verb + EXAMINATION], in which the deverbal noun *examination* possesses its own argument structure, revealed the types of verbs preferred in SVCs and the grammatical properties of *examination* (e.g., the occurrence of an *of*-phrase).

1. Introduction

We typically have the following options when referring to the event of "inspecting someone or something to determine their nature or condition, or testing someone's knowledge or proficiency by requiring them to answer questions or perform tasks" (*New Oxford American Dictionary*, s.v. *examine,* v, 3rd ed.):

(1) a. A doctor *examined* me and said I might need a cesarean.
b. The colleges *examined* candidates.

(2) a. A doctor made an *examination* of the need of a cesarean.
b. The colleges conducted an *examination* of candidates.

From a syntagmatic viewpoint, both examples (1–2) appear to realize the standard transitive verb construction (i.e., [Verb + Obj]). However, in (1), the "examining" event is expressed as the main verb *examine*, whereas in (2), the object noun "examination" (rather than the verb "made" or "conducted") signifies its event. In (2a), the verb *made* taking *examination* as its complement does not literally refer to any actual process of creation (or "making"). Rather, the lexical meaning of the main verb becomes bleached. Traditionally, such verbs are referred to as **light verbs** (Jespersen, 1940; Wierzbicka, 1982) or **support verbs** (Fillmore et al., 2003; Fujii & Uegaki, 2008).

Support verb constructions (SVCs), also known as **light verb constructions** (LVCs)[*1], constitute a subclass of transitive verb constructions (Fujii & Uegaki, 2008). Previous studies have focused on the behavior of typical support verbs (e.g., *do*, *make*, *have*), although there are several explanations that differ in the extent to which verbs contribute to the semantic and syntactic properties of SVCs. This approach, which we refer to as the **verb-centered approach**, focuses on typical support verbs and aims to figure out the process of argument transfer (Grimshaw & Mester, 1988; Grimshaw, 1990), the syntactic and semantic contribution of support verbs (Brugman, 2001; Newman, 1996; Wierzbicka, 1982), and the collocation (Giparaitė, 2023).

However, two issues must be addressed to elucidate the linguistic knowledge that allows speakers to use SVCs. First, "How do we differentiate light senses of verbs from heavy ones?" The verb-centered approach often assumes *a priori* that the verbs in question are support verbs. However, given that verbs used as support verbs can also function as regular transitive verbs, which Brugman (2001) calls using them in a heavy sense, they possess inherent ambiguity. Second, some lexical items function as support verbs only in combination with specific complements (e.g., *sustain an injury*), so when investigating them, there is a risk of excluding verbs that are less frequently used as support verbs.

To address these issues of ambiguity and coverage, verbs should not be described as distinct lexical items but as parts of constructions—conventionalized associations of meaning and form (Goldberg, 2006; Taylor, 2012; Hoffmann, 2022). Support verbs can then be detected by analyzing their complements (cf. Langer, 2005). For a verb to occur in an SVC as a constructional unit, an analysis of semantic properties is needed along with investigation of formal characteristics to differentiate SVCs from regular transitive verb constructions within the configuration of constructions, as in Figure 1.

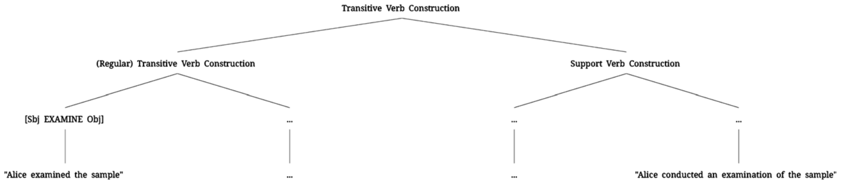

Transitive Verb Construction
(Regular) Transitive Verb Construction
Support Verb Construction
[Sbj EXAMINE Obj]
...
...
...
"Alice examined the sample"
...
...
"Alice conducted an examination of the sample"

Figure 1. The configuration of regular transitive verb constructions and support verb constructions

In this paper, we demonstrate the need for formal specifications of SVCs, focusing on complements, particularly event nouns that possess their own argument structure (e.g, *examination*), to address the issues inherent in the verb-centered approach. Specifically, we try to answer the following questions: (1) Does association strength (collostructional strength) impact the likeliihood of a verb to be considered a light or heavy verb?
(2) Does the presence of an *of*-phrase following the complement exhibit relatively strong predictive power for SVCs?

2. SVCs as a construction

Section 2 provides an overview of previous research on SVCs and points out the methodological and empirical issues inherent in the verb-centered approach. We highlight the need for cues to distinguish verbs with light senses from those with heavy senses, and SVCs from regular transitive verb constructions.

2. 1. Verb-centered approaches to SVCs

In traditional English grammar, support verbs are supposed to lack independent meaning, with their constructions expressing events through the complement (Jespersen, 1940). SVCs are not only observed in English, but widespread among several languages (e.g., Japanese, Korean, German, and Russian). The meaning of the verb phrase relies heavily on the nominal complement, which also determines the argument structure of the verb (cf. Grimshaw & Mester, 1988; Grimshaw, 1990). A representative study of SVCs by Grimshaw and Mester (1988) discussed the Japanese support verb *suru* (trans.: “do”), as in *shuppatsu-o suru* (trans.: “departure”). The authors analyzed *suru* as lacking θ-role assignment capacity and therefore having an incomplete argument structure, proposing the process of argument transfer through which the support verb inherits the argument structure of its complement, enabling it to function within the construction (Grimshaw & Mester, 1988).

Nevertheless, such a process alone fails to explain the preference for certain combinations (e.g., *have a drink*, **have an eat*) and misses the subtle nuances imposed on SVCs but lacking from their regular transitive construction counterparts. Cognitive linguists adopt the gradient view that the light and heavy sense of a particular verb are not strictly categorical (e.g., Brugman, 2001; Newman, 1996; Wierzbicka, 1982). This perspective highlights the fuzzy nature of support verbs, supported by the constraints on several properties support verbs can take such as manner of action, aspect, and valency. For instance, Wierzbicka (1982) pointed out that while it is possible to say "have a drink," meaning *to drink something*, it sounds odd or unacceptable to say "have a study" (or *work*, or *practice*) to express the action in question. Based on those observations, she postulates a prototypical condition under which the support verb *have* is applicable to SVCs, namely as an "AIMLESS OBJECTLESS INDIVIDUAL ACTIVITY WHICH COULD CAUSE ONE TO FEEL GOOD" (Wierzbicka, 1982, p. 762).

In addition, support verbs can impose aspectual constraints on events. For instance, the verb *shower* in (3a) lacks a specific endpoint (i.e., atelic), making it incompatible with adverbial phrases like [*in* + TIME], which indicate the completion of an event within a limited duration of time (i.e., telic). However, Brugman (2001, p. 556) reports that when the verb *shower* is replaced with the SVC *take a shower*, the sentence in question becomes acceptable.

(3) a. Ashley showered { for / $^{?}$in } 10 minutes.
b. Ashley took a shower { $^{\#}$for / in } 10 minutes.

Furthermore, some support verbs retain the valency of their heavy-sense counterparts. Newman (1996) pointed out that SVCs headed by the verb *give* typically demand a dative phrase (*to NP*), just as the heavy sense of *give* does. This observation calls into question Grimshaw and Mester's (1988) view that support verbs do not possess an independent argument structure.

(4) a. $^{\#}$John gave a presentation.
b. John gave a presentation to his students.

In the following discussion, we refer to the methodology applied in the studies reviewed in this section as the “verb-centered approach” because the explanations rely heavily on the properties of support verbs in SVCs.

2.2. SVCs as constructional units

As seen in Section 2.1., verb-centered approaches take the inventory of support verbs as granted. The following problems are inherent in verb-centered approaches: (i) the ambiguity between “light” uses of verbs (e.g., “Alice made an appointment”) and “heavy” uses of verbs (e.g., “Alice made a breakfast”), and (ii) a relatively low coverage of SVCs. Previous studies have not addressed these issues.

Analysts must differentiate between light and heavy sense in context to explicitly describe the linguistic knowledge that allows speakers to use SVCs. It is worth noting that most verbs used as support verbs in SVCs can also appear in regular transitive verb constructions, which Brugman (2001) calls the *heavy* sense.

(5) a. A doctor *made*$_{light}$ an early diagnosis.
 b. A doctor *made*$_{heavy}$ a medical certificate.
(6) a. Alice *sustained*$_{light}$ injury.
 b. Alice *suffered*$_{light}$ a loss.
 c. Alice *wage*$_{light}$ war.

(cf. Fillmore et al., 2002, p. 790)

In (5a), the verb *made* functions as a support verb, inheriting the semantics and argument structure of the complement *diagnosis*. In contrast, in (5b), *made* conveys a heavy sense (literally “make”). Verbs in (6) are examples of lexical items that are typically interpreted as having a heavy sense but peculiarly function as support verbs with a very limited set of complements. When determining whether a given verb in [Verb + Obj] is used as a support verb or a regular transitive verb, and when extending the scope of investigation beyond common support verbs, we must, at the very least, refer to its complement. The verb-centered approach often assumes *a priori* that the verbs in question are support verbs. However, this approach ignores the potential ambiguity of verbs, thus posing the risk of excluding verbs that are less frequently used as support verbs from the scope of investigation. To address this issue, verbs should not be reduced to distinct lexical items but rather described as parts of constructions—conventionalized associations of meaning and form (Goldberg, 2006; Taylor, 2012; Hoffmann, 2022).

As we have seen, Brugman (2001) emphasizes the continuity between support verbs and heavy-sense verbs in terms of semantic contribution,and seems reluctant to establish a clear delimitation between SVCs and regular transitive verb constructions. Interestingly, she makes the following remarks regarding the characterization of SVCs as constructions.

> There are certainly reasons to talk about an LVC (taking even the verb head as variable rather than specified), given the common semantic relationships associated with schema extraction and their consequent properties — we can say with some assurance that, fuzzy as they are, there are some Aktionsart properties common to all LVCs by contrast with their monomorphemic paraphrases. (Brugman, 2001, p. 576)

Following this approach, it is possible to postulate a set of constructions subsumed under the support verb construction, as shown in Figure 2. In the configuration, the subschema (e.g., [*take* + OBJ]) inherits the abstract specification of superschema (SVC) and elaborates the semantic constraints (e.g., aspect, manner of action) that each support verb imposes on their complements. Linguistic knowledge is constructed in a bottom-up fashion and conceptualized as an extensive inventory of actual usage patterns (Taylor, 2012) when adopting the usage-based model (Langacker, 2000), a view that aligns closely with the principles of construction grammar (Hoffmann, 2022).

If speakers' linguistic knowledge of SVCs constitutes an inheritance structure (as in Figure 2), verb-centered approaches only deal with a handful of SVC subclasses. Analyzing SVCs with "major" support verbs could lead to misguided generalizations. To address this issue, analysts should treat the verb slots as variables rather than constants (cf. Uchida, 2010). In this way, they should be able to observe the ambiguity of light and heavy senses of collocating verbs and mediate the coverage.

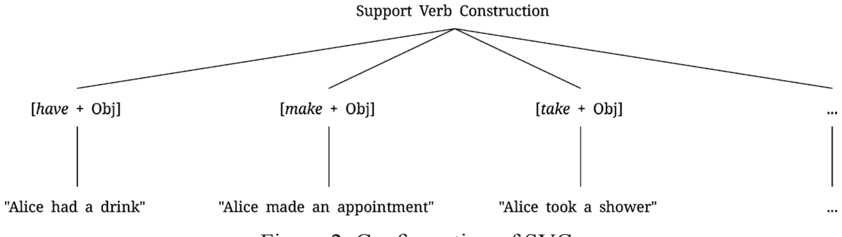


Figure 2. Configuration of SVCs

To support this claim, we extracted instances from a corpus where the deverbal noun examination, which has its own argument structure, functions as the complement of a verb, conducting both quantitative and qualitative analyses of [Verb + EXAMINATION]*2. Through the analyses, we found at least two cues to identify the SVCs in this syntactic environment: (i) verbs with high association strength (collostructional strength) and (ii) an *of*-phrase following EXAMINATION (e.g., "the examination of old age and society").

3. Methods

To observe the formal environment of SVCs, we used Sketch Engine (Kilgarriff et al., 2004, 2014) to extract all instances of the noun *examination* that occurred as an object of a verb. The initial query yielded 1,117 cases. We excluded cases where the noun *examination* was not realized as an object of a verb, resulting in 1,036 cases and a type frequency of 231.

We annotated the instantiation of SVCs based on the realizations of relevant semantic roles*3. Since characterizing SVCs can be challenging, we employed the manageable semantic role-based characterization of SVCs as defined in (7), similar to that employed in FrameNet (Fillmore et al., 2002, 2003). We also defined EXAMINING as in (8) to annotate the distribution of *examination*-related semantic roles.

(7) Semantic role-based characterization of SVCs: The construction containing *examination* is an instance of SVCs if and only if the noun in the subject position realizes at least one semantic role of EXAMINING regardless of the collocating verbs (cf., "Alice {passed, conducted} an examination").

(8) EXAMINING: <Examiner> assesses the <Attribute> of <Examinee>

a. [$_{\text{<Examinee>}}$ Alice] **passed** an examination.

b. [$_{\text{<Examiner>}}$ Alice] **conducted** an examination.

Then, for each case, we annotated the following formal features to identify the formal environment that SVCs prefer. First, we annotated the voice of the construction containing *examination* as is_passive (9a). Most transitive constructions can be realized in passive or active voice, making them candidates for the crucial formal SVC environment. We also annotated the realization forms of noun phrases by coding the grammatical number as noun_is_singular (9b) and the presence of any article (9c).

(9) Formal environments of SVCs:

a. **is_passive:** 1 iff the case in question is realized in passive voice, 0 otherwise.

b. **noun_is_singular:** 1 iff the noun *examination* is realized as a singular noun, 0 otherwise.

c. **has_article:** 1 iff the noun *examination* co-occurs with an (indefinite or definite) article, 0 otherwise.

d. **coll_strength:** Pearson residuals between expected and observed frequency of verbs.

To compute the collocational strength, we performed collostructional analysis using Coll.analysis 4.1. (Gries, 2024). The term "collostructional analysis" refers to a family of collocational analyses that can accurately capture the collocational strength between grammatical constructions and words (Stefanowitsch & Gries, 2003, 2005; Gries & Stefanowitsch, 2004a, b; Gries, 2019, 2023). While many association measures are currently available in corpus linguistics, employing Pearson residuals was proposed to measure the degrees of collocational preference (Gries, 2023). Pearson residuals refer to the difference between observed and expected frequency in the form of a cross-tabulation table. Using this approach, analysts can capture the words' preference (or repulsion) in the construction. Pearson residuals were employed to determine collostructional strength.

We performed logistic regression analysis (Gries, 2021; Levshina, 2015; Speelman, 2014) to explore the extent that predictors in (9) contribute to discriminating SVCs from non-SVCs. Logistic regression analysis is a type of regression analysis using categorical response variables (i.e., every sentence in question is either an SVC or not). Regression analysis can reveal differences in the data and predict the variables contributing to the distribution of response variables. Performing logistic regression analysis allows analysts to determine the contribution of predictors in classifying constructions.

Moreover, regression analysis provides a formula that predicts the distribution of its response variable (i.e., isSVC). This allows analysts to compute the extent of correct data predictions using a confusion matrix consisting of the frequency of predicted and actual instances. For instance, the contingency table shown in Table 1 reveals 20 misclassified items (10 false positives and 10 false negatives). Analysis of a confusion matrix can be used to evaluate the performance of a constructed model.

All annotations were carried out by the authors. We used R (R Core Team, 2024) to perform the computation and a family of ggplot2 to visualize the results (Wickam et al., 2024).

Table 1. A fictitious distribution of predicted and actual frequency of isSVC

	isSVC (Predicted)	¬ isSVC (Predicted)
isSVC (Actual)	90	10
¬ isSVC (Actual)	10	90

4. Results

This section reports the result of our corpus study. As a result, we revealed that the interpretation of SVC is likely to be realized when the verb slot of SVC is filled with verbs with high association strength. We report quantitative and qualitative results of our study. All the codes and data used in this study are available on the Open Science Framework (OSS).

4.1. Descriptive statistics

4.1.1. Types of verbs

Of 231 verb types, 99 were realized as SVCs and 164 types were not. While the realization of SVC was not mutually exclusive in some verbs (e.g., allow), token

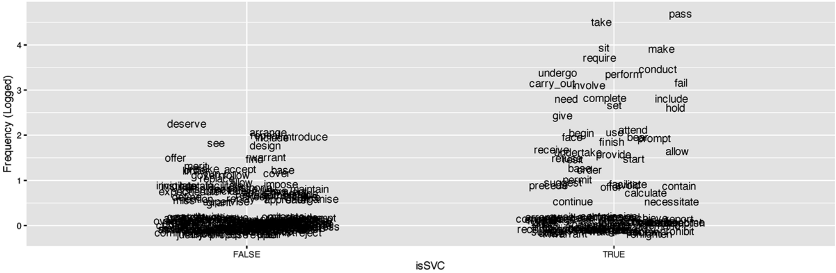


Figure 3. Distribution of raw frequency and isSVC

frequency of SVC was higher than the transitive uses (759 cases realized as SVC and 277 cases realized as non-SVC). Figure 3 shows the distribution of logged frequency on the y-axis and the types of constructions in which verbs occur.

Given that the raw frequencies do not necessarily convey accurate association (Kambara & Chika, 2023; Kambara et al., in press), the association strength of verbs and EXAMINATION was computed using Pearson residuals as collostructional strength. Figure 4 shows the distribution of Pearson residuals in SVC and non-SVC constructions with the boxplot under each collocate. The height of the boxplot in Figure 4 suggests that the distribution of Pearson residuals is spread more widely in non-SVCs because that type includes low frequency verbs. In contrast, values of Pearson residuals in SVCs show a relatively more even spread across the y-axis. This result suggests that a high association strength between the noun and verbs is a strong cue for identifying an SVC.

We can deduce that lexical items functioning as support verbs, occurring in SVCs, tend to exhibit relatively high association strength. Additionally, verbs used as heavy sense, occurring in regular transitive verb constructions, generally show low collostructional strength with EXAMINATION, except for a few outliers.

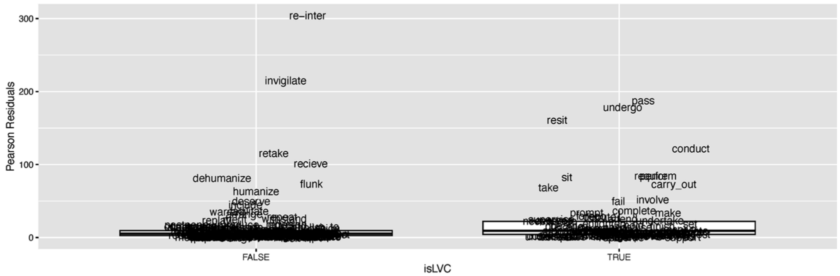


Figure 4. The distribution of Pearson residuals in SVCs

4.1.2. Grammatical features of EXAMINATION

In addition to the types of verbs, we also recorded the morpho-syntactic environment of the noun *examination*, focusing on three variables: has_article, noun_is_singular, and collocates_with_of. The raw frequencies of these three variables are summarized as cross-tabulations in Table 2 and visualized as three distinctive mosaic plots in Figure 5.

Table 2. Raw frequency of each morpho-syntactic variable

	isSVC	¬ isSVC
has_article	559	176
¬has_article	200	101
noun_is_singular	533	146
¬noun_is_singular	226	131
collocates_with_of	175	1
¬collocates_with_of	584	276

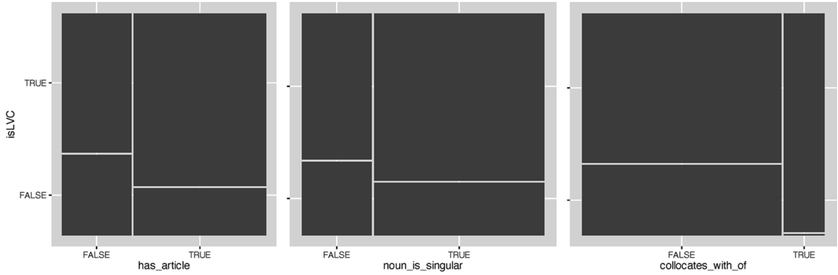


Figure 5. Morpho-syntactic environments of *examination*

In the mosaic plots, the widths of the bars represent the proportional distribution of the variable on the x-axis and within each of the (stacked) bars, the heights indicate the proportional distribution of the levels of the variable on the y-axis (Gries 2021, p. 123). The plot indicates two types of features that most instances of *examination* are likely to realize: (i) as singular and with an article, and (ii) the presence of an *of*-phrase as shown in (10).

(10) a. […] his research team began the electrical examination of acupuncture points of human system […] [CB9 1436]

b. This project conducts an examination of old age and society between 1918 and 1948 […] (HJ0 3670)

4.2. Inferential statistics

Based on the descriptive statistics, we constructed a statistical model using a logistic regression analysis using collocates_with_of and collostructional strength. The effects of each predictor are shown in Figure 6. The constructed binomial model, represented in Table 3, was statistically significant with each of the predictors significantly contributing to the discrimination of the presence of SVC. The plot and table show the presence of the preposition *of* and collostructional strength contribute to the realization of SVCs.

A notable finding from Table 3 is that combinations where an *of*-phrase follows almost certainly instantiate SVCs. In contrast, at least within the scope of the features annotated in this paper, other grammatical features were not interpreted as crucial factors in determining whether [Verb + EXAMINATION] instantiates SVCs or regular transitive verb constructions, that is, whether the verb in question is a support verb or not.

Table 4 shows the confusion matrix of isSVC. The accuracy of the model was 83.5%. Given that the SVC proportion was high in the observed data, we opted to set the baseline by computing the proportion of maximum value, which was 74% (= 0.7326). The accuracy exceeded the baseline, and we concluded that the model made a fairly "good" classification. Nagelkerke's Pseudo R^2 was calculated at 0.548, suggesting that the constructed model was partly effective. The C score was 0.917, showing outstanding discrimination of the model (Levshina, 2015, p. 256). These results suggest that SVC classification is a fairly easy task when collostructional strength is taken into account, and that the presence of the preposition *of* can help

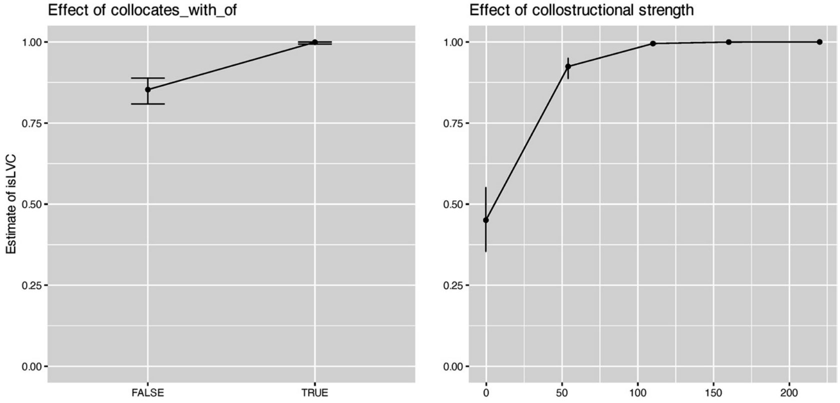


Figure 6. Effect plot of formal predictors

listeners readily identify the construction.

Table 3. Coefficient table of the constructed model

	Estimate	Std. Error	z value	Pr (> |z|)	
(Intercept)	−1.077	0.132	−8.1250	0.000	***
collocates_with_of TRUE	5.248	1.009	5.200	0.000	***
coll_strength	0.050	0.004	12.525	0.000	***

Table 4. Confusion matrix of isSVC

	isSVC (Predicted)	¬ isSVC (Predicted)
isSVC (Actual)	647	112
¬ isSVC (Actual)	59	218

5. Discussion

The overall findings of the result in Section 4 can be summarized as follows:

(11) Overall findings:

a. When verbs with high collostructional strength collocate with EXAMINATION, the verb phrase [Verb + EXAMINATION] generally instantiates SVCs (Figure 4).

b. Among the grammatical properties of *examination*, the occurrence of an *of*-phrase is the strongest predictor for [Verb + EXAMINATION] instantiating SVCs (the left panel of Figure 6).

c. Based on findings (11a-b), it is relatively straightforward to predict which constructions [Verb + EXAMINATION] instantiate (the right panel of Figure 6, Table 4).

In this section, we discuss the qualitative results and the implications of describing SVCs as a type of constructional unit, based on our findings.

Regarding (11a), it is possible that verbs with high collostructional strength form a cluster, sharing certain semantic features and providing information such as [Verb$_i$ + EXAMINATION] ↔ <examining in the manner/aspect specified by *i*>*4. Given that SVCs are commonly used in everyday language, they might appear to exhibit high productivity (i.e., creativity). However, our findings suggest the opposite—that their productivity is relatively low. In other words, this SVC applies to a relatively limited group of verbs.

Figure 7 displays the ten verbs with the highest and lowest association strengths with SVCs. Some verbs in Figure 7 have not been previously described in SVC studies (e.g., *resit, conduct,* and *pass* in [Verb + EXAMINATION]). While this finding supports our claim in Section 2.2 that the verb-centered approach alone results in low coverage, many verbs with low collostructional strength (e.g., *commission, prove, achieve*) were also found in SVCs. Thus, caution is necessary when predicting that verbs with low collostructional strength with EXAMINATION instantiate regular transitive verb constructions rather than SVCs.

In (11b), the complement following a support verb is generally characterized as possessing an argument structure and functioning as an input for argument transfer (Grimshaw, 1990), as reviewed in Section 2.1. When distinguishing SVCs from regular transitive verb constructions within a constructional network, as represented in Figure 1, it may be effective to specify the slot for the argument instantiated as an *of*-phrase. However, while this specification suggests that the presence of an *of*-phrase serves as a sufficient condition for SVCs, the absence of *an* of-phrase does not necessarily imply that the given environment is not an SVC. Based on the frame-semantic identification criteria presented in (7–8), an *of*-phrase appears to fulfill one of the frame elements within the EXAMINING frame (e.g., <Examiner>, <Examinee>). Therefore, to further examine the relationship between the presence of an of-phrase and whether the given environment instantiates SVCs, additional analysis focusing on the frame elements within the *of*-phrase is necessary.

Finally, as a key finding from the logistic regression analysis conducted in Section 4.2, (11c) indicates that the features annotated in this study provide a relatively accurate model for determining whether [Verb + EXAMINATION] instantiates SVCs. However, as discussed in relation to (11a), this predictive power is likely offset by the idiomaticity of the construction, suggesting that the constructional knowledge enabling the use of SVCs may be more specific than the sketch presented in Figure 2. Future

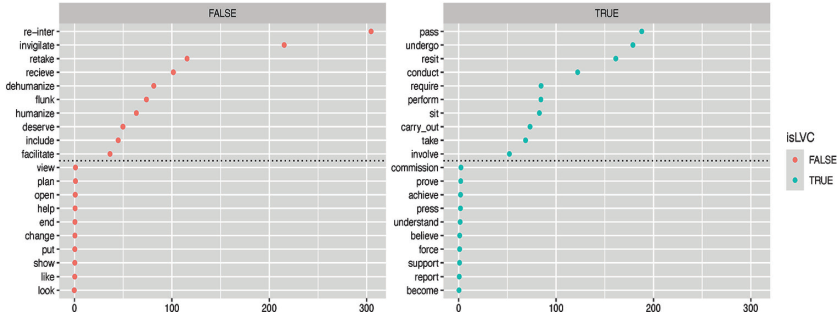


Figure 7. Verbs with the highest and lowest association strength with SVCs

research should explore the appropriate level of abstraction for constructional units.

6. Conclusion

This paper presented a corpus-based survey of [Verb + EXAMINATION] to explore the formal specifications of SVCs. To describe the linguistic knowledge that enables speakers to use SVCs, we must address two issues: the contextual ambiguity between support verbs and their heavy-sense counterparts, and the limited coverage of low-frequency support verbs. We argue for the need for formal specifications of complements to address the limitations inherent in the verb-centered approach.

Our results in Section 4 showed (i) verbs with high collostructional strength, and (ii) the grammatical properties of EXAMINATION (e.g., the occurrence of an *of*-phrase) when [Verb + EXAMINATION] instantiates an SVC. Furthermore, we found that verbs such as *resit, conduct,* and *pass*, which exhibit high collostructional strength with *examination*, can function as support verbs—types of verbs that have not been addressed in previous studies on SVCs. These findings indicate the importance of examining formal properties of complements in addition to semantic constraints when describing SVCs as constructions, which Brugman (2001) emphasized in relation to aspectual constraints.

Another limitation of this study is that the scope of analysis was restricted to EXAMINATION. First, it is necessary to examine whether the deverbal noun EXAMINATION can represent the behavior of SVCs where deverbal nouns of the V-*tion* type function as complements. Additionally, further investigation is required to identify the formal characteristics of SVCs when deverbal nouns other than the *-tion* type (e.g., *make a catch, take a test*) occur as complements.

Notes

[*1] According to Fujii and Uegaki (2008), SVCs and LVCs share many semantic and syntactic properties. These authors distinguish between LVCs and SVCs from the perspectives of Frame Semantics and Construction Grammar. Specifically, they classify constructions like *make a complaint*, where the verb's semantic contribution is minimal and the construction exhibits high generality, as LVCs. In contrast, constructions like *lodge a complaint*, where the verb contributes to the overall meaning of the phrase and displays a higher degree of idiomaticity, are categorized as SVCs. In the following discussion, the broader term SVCs and support verbs will be used.

[*2] In the following discussion, EXAMINATION in [Verb + EXAMINATION] represents any NPs headed by *examination*.

[*3] We also considered the results of Kambara (2021) during the annotation process. While Kambara (2021, p. 154) presents 28 verbs used as LVCs, we included all verbs extracted from the corpus.

[*4] To represent the constructional schema, we divide [Verb + EXAMINATION] into a phonological pole and a semantic one. The index *i* indicates the correspondence relation between the phonological and semantic poles.

Acknowledgements

Portions of this work were supported by JSPS KAKENHI Grant Number 24K16143. We thank the comments from the two anonymous reviewers. Usual disclaimers apply.

Data availability

All codes and data are available on OSF (Open Science Framework): 10.17605/OSF.IO/4TAEK

Dictionaries

New Oxford American Dictionary. 3rd ed. 2010.

References

Fillmore, C. J., Baker, C. F., & Sato, H. (2002). Seeing arguments through transparent structures. In *Proceedings of Third International Conference on Language Resources and Evaluation* (*LREC 2002*) (pp. 787–791).

Fillmore, C. J., Johnson, C. R. & Petruck, M. R. L. (2003). Background to FrameNet. *International Journal of Lexicography*, *16*(3), 235–250.

Fujii, S. & Uegaki, W. (2008). Shien doushi koubun no bunseki: Ko-pasu ni motoduku koubun riron teki apuro-chi [An analysis of Support Verb Constructions: A constructionist approach based on corpus]. *Proceedings of the thirteenth annual meeting of the association for natural language processing* (pp. 943–946) [Written in Japanese].

Giparaitė, J. (2024). A corpus-based analysis of light verb constructions with MAKE and DO in British English. *Kalbotyra*, *76*, 18–41. https://doi.org/10.15388/Kalbotyra.2023.76.2

Goldberg, A. E. (2006). *Constructions at works: The nature of generalization in language*. Oxford University Press.

Gries, S. T. (2019). 15 years of collostructions: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics*, *24*(3), 385–412.

Gries, S. T. (2021). *Statistics for linguists with R: A practical introduction*. De Gruyter Mouton.

Gries, S. T. (2023). Overhauling collostructional analysis: Towards more descriptive simplicity and more explanatory adequacy. *Cognitive Semantics*, *9*(3), 351–386.

Gries, S. T. (2024a). *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. John Benjamins.

Gries, S. T. (2024b). *Coll.analysis 4.1. A script for R to compute and perform collostructional analyses*. https://www.stgries.info/teaching/groningen/index.html.

Gries, S. T. & Stefanowitsch, A. (2004a). Co-varying collexemes in the into-causative. In M. Achard & S. Kemmer (Eds.), *Language, culture, and mind* (pp. 225–236). CSLI.

Gries, S. T. & Stefanowitsch, A. (2004b). Extending collostructional analysis: A corpus-based perspective on alternations. *International Journal of Corpus Linguistics*, *9*(1), 97–129.

Grimshaw, J. (1990). *Argument structure*. MIT Press.

Grimshaw, J. & Mester, A. (1988). Light verbs and θ-marking. Linguistic Inquiry, 19, 205–232.

Hoffmann, T. (2022*). Construction grammar: The structure of English*. Cambridge University Press.

Jespersen, O. (1940). *A modern English grammar on historical principles: Part V., Syntax*.

Copenhagen.

Kambara, K. (2021). *Fure-mu imiron ni motoduku meishi no imi bunseki* [Frame semantics analysis of nominal meanings] [Doctoral dissertation, Graduate School of Human and Environmental Studies, Kyoto University]. https://doi.org/10.14989/doctor.k23581.

Kambara, K. & Chika, T. (2023). Toward a corpus-based identification of nominal relationality and uniqueness: A constructionist approach. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation* (pp. 661–669).

Kambara, K., Takahashi, N., & Chika, T. (in press). Quantifying the degrees of relationality: A collostructional approach. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The sketch engine. In G. Williams and S. Vessierssier (eds.), *Proceedings of 11th EURALEX International Congress* (pp. 105–116).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., ... & Suchomel, V. (2014). The sketch engine: Ten years on. *Lexicography*, *1*(1), 7–36.

Langacker, R. W. (2000). A dynamic usage-based model. In *Grammar and conceptualization* (pp. 1–63). De Gruyter Mouton.

Langer, S. (2005). A formal specification of support verb constructions. In Langer, S. & Schnorbusch, D. *Semantik im lexikon* (pp. 179–202). Narr Dr. Gunter.

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins.

Newman, J. (1996). *GIVE: A cognitive linguistic study*. John Benjamins.

R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Stefanowitsch, A. & Gries, S. T. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209–243.

Stefanowitsch, A. & Gries, S. T. (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory*, *1*(1), 1–43.

Speelman, D. (2014). Logistic regression: A confirmatory technique for comparisons in corpus linguistics. In D. Glynn and J. A. Robinson (eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (pp. 487–533). John Benjamins.

Taylor, J. R. (2012). *The mental corpus: How language is represented in the mind*. Oxford University Publishing.

Uchida, S. (2010). On the lexicographic descriptions of event nouns: An insight from frame semantics. *The Journal of Institute for Language and Education Research*, *27,* 411–426.

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media.

Wierzbicka, A. (1982). Why can you *have a drink* when you can't **have an eat* ?. *Language* 58 (4), 753–799.

（近　大志　京都大学）
（神原　一帆　情報通信研究機構／立命館大学）

「論文」

Developing the preliminary essay bundles list (EBL) and its applicability to EAP

Ryo SAWAGUCHI

Abstract

This study developed the preliminary list of lexical bundles (e.g., *on the other hand*, *the fact that*) for argumentative essay writing and explored its potential applications to English for General Academic Purposes (EGAP) practice to prepare undergraduate students for their future use in academic written English genre (e.g., research papers). The list, called the Essay Bundles List (EBL), was created by extracting frequently used lexical bundles from opinion- and source-based argumentative essays by L1 English speakers. Corpora consulted include the International Corpus Network of Asian Learners of English (ICNALE), Louvain Corpus of Native English Essays (LOCNESS), PERSUADE2.0, Michigan Corpus of Upper-Level Student Papers (MICUSP), and British Academic Written English (BAWE). A total of 3,768 bundles were compared with the list of academic written English (Academic Formulas List: AFL) to confirm EBL applicability. The results showed that the EBL covers approximately 80% of the AFL, indicating its potential as an EGAP wordlist. Correspondence analysis of the top 21 frequent bundles in opinion- and source-based essays and the AFL revealed that the opinion-based bundles (e.g., *I believe that*) can be made suitable for academic written English with the use of inanimate subjects (e.g., *it is true that*), while source-based discourse bundles (e.g., *in order to*) imply their direct applicability. The EBL was refined to 127 bundles according to their difficulty levels on the Common European Framework of Reference (CEFR) scale (A2, B1, B2) in proficiency order. This study suggested that, basic referential bundles (e.g., *the fact that)* and objective stance bundles (e.g., *this means that*) are appropriate for A2 and B1 students. Discourse bundles (e.g., *to begin with*, *on the one hand*) should remain a focus throughout the progression from B1 to B2. Advanced referential bundles, such as *the existence of*, are most suitable for instruction at the B2 level.

Introduction

Corpus linguistics developments in the last few decades have made it possible to analyze recurring word clusters called lexical bundles (e.g., *in order to*, *as a result of)*. They have been examined in English for Academic Purposes (EAP) contexts (Biber et al., 2004; Biber & Barbieri, 2007; Hyland, 2008) because of their significant discourse functions in academic speech or writing. Lexical bundles are “multiword sequences that occur most commonly in a given register” (Biber & Barbieri, 2007, p. 264). Register refers to variations of language (e.g., spoken or written) used in different situational characteristics. According to Biber et al. (2004), classroom conversations are regarded as an *oral* register, while academic prose is defined as a *literate* register. Biber and Barbieri (2007, p. 273) comment that “the extent to which a speaker or a writer relies on lexical bundles is strongly influenced by their communicative purposes.” For example, the bundle *it is clear that* expresses a writer’s point of view in the following sentence, while *as a result* connects the preceding and subsequent sentences. The significance of lexical bundles has resulted in the development of a useful wordlist, the Academic Formulas List (AFL: Simpson-Vlach & Ellis, 2010), to assist in the intensive learning of lexical bundles (which researchers term *formulas*). However, no list of lexical bundles for college-level writing genres such as essays has yet been developed.

To prepare students for future academic English situations (e.g., writing papers), argumentative essays have been the most common writing genre for undergraduate students (Wu, 2006). The possible reasons could be the applicability of argumentative essays to research papers in terms of genre and text types. Swales (1990) defines genre as a set of events sharing the same communicative purposes. Biber (1989) describes text types as differences in linguistic features. In case of argumentative essays and research papers, the two express the writer’s stance and support it with evidence; thus, these can be classified under the argumentative writing genre with the same communicative purpose: arguing. Johnson (2018) claims that the rhetorical characteristics in genres such as argumentation and exposition encompass various text types. Consequently, argumentative essays and research papers possibly share the similar linguistic features, e.g., the same lexical bundles. Argumentative essays have also been used to assess the use of lexical bundles by undergraduate students to improve their basic academic writing skills (Granger, 2017; Nam & Park, 2020; Sawaguchi, 2024), a foundation for English for General Academic Purposes (EGAP; Blue, 1988) in which "students from a wide range of disciplines will write in diverse genres" (Tardy et al., 2022, p. 3). Despite the significance of argumentative essay writing in EGAP programs, no lexical bundle wordlist has yet been developed for intensive vocabulary learning for essay writing. In EAP contexts, wordlist use has facilitated greater student academic vocabulary use (Shoufaki & Petrić, 2021). Given the effectiveness of the EAP wordlist, there is a vital need for an argumentative essay writing wordlist. Another unexplored area of argumentative essay writing is its applicability to undergraduate students' future use of academic written English genre (e.g., research papers).

Therefore, developing an argumentative essay lexical bundle wordlist for academic English could significantly encourage more meaningful and focused EGAP writing practice. This study develops a possible lexical bundle list for argumentative essay writing and investigates its relevancy to academic written English. Practical suggestions on the use of the list are also proposed.

Literature Review

Essays are an important text type of writing in higher education settings (Nesi et al., 2017), with argumentation in particular often being a key student requirement (Wingate, 2012) for the development of critical/logical thinking and rational argument skills.

There has been significant research into the pedagogical applications of argumentative essays. To gain a more precise understanding of argumentative essays, Yoon and Tabari (2023) classified argumentative essays into two categories: source-based and opinion-based. In source-based essays, writers organize and present their arguments based on established information sources (e.g., research articles). By contrast, opinion-based essays require the writer's knowledge or experience: the topics include the pros and cons of part-time jobs for university students.

Despite the noted importance of lexical bundles in academic English, few studies have investigated how these lexical bundle items are dealt with in argumentative essay teaching materials or curricula. Sawaguchi (2024) focuses on identifying target lexical bundles for opinion-based argumentative essay writing using the L1 English speaker essay corpora in the Louvain Corpus of Native English Essays (LOCNESS: Granger, 1998) and the International Corpus Network of Asian Learners of English (ICNALE: Ishikawa, 2023), and identifies the most frequent lexical bundles in various opinion-based essay topics. He proposed a teaching order for the target bundles based on the estimated difficulty for Japanese university students, with a particular focus on the bundles they are unfamiliar with.

However, EAP applications must include target lexical bundles for both opinion- and source-based essays, as both argumentative writing types are included in EGAP writing courses. It would also be valuable to clarify the lexical bundle differences between these two essay types. While both share the common moniker of "argumentative essays," it is likely that the lexical bundles vary because of their different argument bases: opinion or source. Understanding the lexical bundle differences for these two essay types could assist EAP teachers in teaching according to their needs. Further, because EGAP writing equips students with general academic writing skills before they proceed to English for Specific Academic Purposes (ESAP: specializing in their own disciplines; Blue, 1988), identifying the associations between argumentative essay lexical bundles and those in academic written English could also be useful for teaching practice. For example, a lexical bundle *on the other hand* frequently occurs in four different disciplines, namely biology, electrical engineering, applied linguistics, and business studies (Hyland, 2008). Given its cross-disciplinary usage, prioritizing *on the other hand* in EGAP writing courses for first- or second-year university students would help them build transferable writing skills that remain useful regardless of their specific discipline when they advance to ESAP contexts (e.g., graduate school studies) and academic writing in their respective fields. This is pertinent to "the nature of a "common core" of features relevant to all types of academic writing, applicable in a wide range of EAP teaching contexts" (Gardner et al., 2018, p. 647), which could allow students to apply their lexical bundle knowledge learned in EGAP to ESAP. The teaching practice would also be more effective if the target lexical bundles have difficulty levels suitable for students at different proficiency levels. Accordingly, this study addressed the following research questions:

RQ 1: To what extent are argumentative essay lexical bundles relevant to those in academic written English?

RQ 2: How can the argumentative writing lexical bundles be categorized according to their difficulty levels?

The study then explored how the findings in RQ 1 and 2 could be applied to EGAP practice.

Data and Procedure

Opinion-Based Essay Corpus

This study used three opinion-based argumentative essay corpora: the ICNALE Written Essays (Ishikawa, 2023), the Louvain Corpus of Native English Essays (LOCNESS) (Granger, 1998), and the PERSUADE2.0 (Crossley et al., 2024). These three corpora include various topics, such as the pros and cons of animal testing and part-time jobs for university students, which are generally based on the writers' own opinions or ideas and do not typically require research-based evidence; therefore, in this paper, I termed these types of argumentative essays "opinion-based essays." The ICNALE, the LOCNESS, and the PERSUADE2.0 were chosen for the following three reasons. First, to the best of the author's knowledge, they are publicly available free corpora containing the essays by L1 English speakers, making it easier for other researchers to replicate the results of the study. Second, the corpora contain target-like lexical bundles, such as A-level essays and those written by L1 English instructors or professors. Third, the corpora have over 20 different essay topics, which allows for the extraction of commonly used lexical bundles in the corpora regardless of the topic. As in Nation's (2016) discussion on the creation of wordlists, range (See *range* in lexical bundle definition and extraction for details) is one of the most important criteria, as useful words should be found in a variety of texts.

During the extraction process, I excluded essays in the LOCNESS that were not argumentative, such as literary and exam essays in the file USMIXED. Table 1 presents the topics, the number of words and files for the opinion-based essay corpus, and the corpora analyzed in the study. The files part-time jobs and smoking in restaurants are from the ICNALE (the pros and cons of part-time jobs and smoking in restaurants). The topics in the LOCNESS were manually categorized into four major topics: human rights (e.g., gender equality), technology (e.g., the invention of computers), politics (e.g., parliamentary systems), and others (e.g., sports, the media). The files seeking opinions (seeking multiple opinions from others) and summer projects (should summer projects be designed by students?) are from the PESUADE2.0.

Table 1. Breakdown of the opinion-based essay corpus in the study

Topics	No. of files	No. of words
Human rights	45	45,835
Politics	55	41,518
Part-time jobs	200	45,415
Seeking opinions	74	45,182
Smoking in restaurants	200	45,198
Summer projects	79	45,028
Technology	118	63,720
Other	120	88,194
Total	891	420,090

Source-Based Essay Corpus

For the source-based essay data, I consulted the British Academic Written English (BAWE; Nesi et al., 2008), which includes course assignment essays from British university students, and the Michigan Corpus of Upper-Level Student Papers (MICUSP; Römer & O'Donnell, 2011), which contains approximately 830 A-grade papers from various disciplines (humanities and arts, social sciences, physical sciences) from the University of Michigan. Because the MICUSP and the BAWE both include research-based essays written by university students from different disciplines, these are defined as "source-based essays" in this study. As with the opinion-based essay data, only essays written by L1 English speakers were extracted. The BAWE and the MICUSP were prioritized over other similar type of corpus: Academic Writing at Ackland (AWA) due to the two corpora's potential large number of words for this study; the BAWE: approximately 580, 000 words; the MICUSP: approximately 450,000 words. These sizes of words were considered adequate for obtaining data from varied disciplines. While AWA was also a potential candidate, integrating it would have required a major adjustment to the data balance in this study. To maintain equal representation across disciplines, the study focused on gathering an equal number of words from the arts and humanities (including social sciences) and the sciences. Given the aim of developing an EGAP essay lexical bundle list applicable across disciplines, the dataset was structured to ensure balanced discipline coverage. Table 2 shows the discipline genres and the total number of files and words analyzed in this study. The arts and humanities were subdivided into specific disciplines, such as archeology, linguistics, and history, and the social sciences had disciplines including business, economics, and education. Compared to the arts and humanities and social sciences, both the BAWE and the MICUSP have a relatively limited number of essays from the life/physical sciences (e.g., biology, physics). Therefore, life and physical sciences were integrated into the sciences to balance the number of words reviewed in both the arts and humanities and sciences to approximately 340,000 words each.

Table 2. Breakdown of the source-based essay corpus in the study

Disciplines	No. of files	No. of words
Arts and humanities	117	348,379
Social sciences	136	345,859
Sciences (life/physical)	171	344,343
Total	424	1,038,581

Lexical Bundle Definition and Extraction

This study defines lexical bundles as three-to-five-word clusters that satisfy the following frequency and range criteria. The reason I focused on three to five clusters is discussed first.

Word cluster length: The RQ 1 of this study is focused on exploring how the lexical bundles in the essay wordlist could be applied to academic written English. To do this, I examined the coverage of the essay wordlist in the AFL, for which I decided the lexical bundle lengths should be the same. For example, the bundle *at the end of the day* is a six-word bundle found in the essay wordlist; however, the AFL limits bundle lengths to five, which results in bundles such as *the end of the*. By setting an equal length for the word clusters, this study sought to discover the bundles that overlap the argumentative essay bundles and the AFL.

Frequency: Lexical bundles occur at least 20–40 times per million or more in different texts (Biber & Barbieri, 2007; Hyland, 2008). These frequency criteria indicate that lexical bundles do not occur by chance but are a representation of linguistic phenomena. This study employed the standard 20 times per million criteria for the extraction of both the opinion- and source-based lexical bundles. Compared to studies such as Hyland (2008), which analyzed 3 million words, this study used relatively small-sized corpora (approximately 1.45 million words in total), primarily because the study sought to identify the frequently appearing lexical bundles in smaller L1 corpora by establishing a minimum frequency threshold.

Range: Range is the extent to which lexical bundles are distributed across various texts. Research has indicated that lexical bundles appear in five or more texts (subcorpora created from the main corpus) (Biber et al., 2004; Bychkovska & Lee, 2017; Omidian et al., 2018). Range is a key criterion for filtering an individual writer's idiosyncratic language use. For example, if one writer uses the bundle *as a result* three times, this would affect the total number of raw frequencies; however, analyzing texts by different writers reduces this risk. I applied different range criteria for the opinion-based and source-based lexical bundle extractions. For the opinion-based bundles, I set a minimum of three different texts because the size of the opinion-based essay corpora in this study was similar to that consulted in Chen and Baker (2016), who set three ranges and analyzed under 1 million words (approximately 200,000 words). By setting a lenient range criterion, this study gathered as many lexical bundles as possible from the relatively small-sized opinion-based essay corpora. For the source-based lexical bundles, I applied five different text criteria because this was the standard criteria in previous studies; Omidian et al (2018), whose corpus size was very similar to this study (1030,000 words), used a five-range criterion.

All extraction processes were performed using the N-gram function in the computer concordance software AntConc Ver. 4.2.4 (Anthony, 2023). The extraction resulted in 3,768 opinion- and source-based lexical bundles. However, among the bundles that met the aforementioned frequency, range, word length criteria but are strongly topic-related bundles such as *part-time jobs* in the topic part-time jobs for university students were manually excluded from the analysis because of their low pedagogical value for essay writing. Specifically, *part-time jobs* had the highest frequency (943 times per million words) followed by *be able to* (567 times per million words) in the three-word bundles in opinion-based corpus. Despite the high frequency of *part-time jobs*, the bundle was excluded from the analysis involving frequency information. In contrast, the bundles consisting solely of function words (e.g., *this is a*) were included for the analysis in accordance with the criteria employed in the AFL, which regards these as lexical bundles. Hereafter, the bundles list is called the Essay

Bundles List (EBL).

Compatible Academic Written£ English

For comparison purposes, this study termed the AFL (Simpson-Vlach & Ellis, 2010) as "academic written English". The AFL is the largest wordlist to date that contains academic English lexical bundles (e.g., *on the other hand*, *as a result of*) commonly used across disciplines (e.g., social sciences, humanities, medicine) whose coverage facilitated comparison with this study's aim to develop an EGAP wordlist that could be applicable regardless of disciplines. Another advantage of the AFL is that it categorizes the lexical bundles into three major functional categories: referential (e.g., *in the case of*), stance (e.g., *it is important to*), and discourse (e.g., *in order to*), which allowed for in-depth interpretations of the similarities and differences between the essay lexical bundles in the study and those in the AFL in terms of discourse functions.

The AFL has both spoken and written academic lexical bundle lists termed as written/spoken AFL respectively. Written AFL consists of the lexical bundles frequent in academic written English text types (e.g., research papers, textbooks), while spoken AFL includes the frequent bundles in spoken academic English registers (e.g., lectures, seminars). The AFL integrates these bundles to the *core* AFL, whose lexical bundles are commonly used in both academic speech and writing. Since opinion-based lexical bundles are often more colloquial (Chen & Baker, 2016), this study chose the core AFL to better assess its coverage in the EBL. Additionally, the core AFL contains more lexical bundles (207) compared to the spoken and written AFL (200 bundles each), making it more extensive for the coverage assessment of the study, which involves a total of 3,768 lexical bundles in the EBL. While the core AFL provides frequency information for both spoken and written academic English, this study focused on the frequency data for written academic English to ensure a consistent comparison with the EBL. Hereafter, the core AFL will be simply termed as AFL.

Results and Discussion

The Applicability of the EBL to Academic Written English

RQ 1 of the study explored the applicability of the EBL to academic written English. For this purpose, the coverage (matching rate) of the EBL and the core AFL lexical bundles was investigated. Table 3 shows the EBL coverage in the AFL and reveals that all lexical bundles in the EBL overlapped 78.7 % of the total 207 lexical bundles in the AFL, which indicates that the EBL has a high degree of coverage in academic written English, and significant potential for inclusion in an EGAP wordlist to prepare students for the future use of academic written English.

Table 3. Coverage of the EBL in the AFL

EBL overlapping bundles	AFL bundles	Coverage (%)
163	207	78.7%

To further explore the frequency relationship between the argumentative (opinion- and source-based) bundles and the AFL (written academic English), a correspondence analysis was conducted on the top 21 frequent AFL (top 10% of the 207 AFL) and the EBL (source/opinion) corresponding 21 bundles using the langtest.jp (Mizumoto, 2015), which is a multifunctional application website that performs statistical analyses. Figure 1 shows the biplot of the correspondence analysis. Dimen-

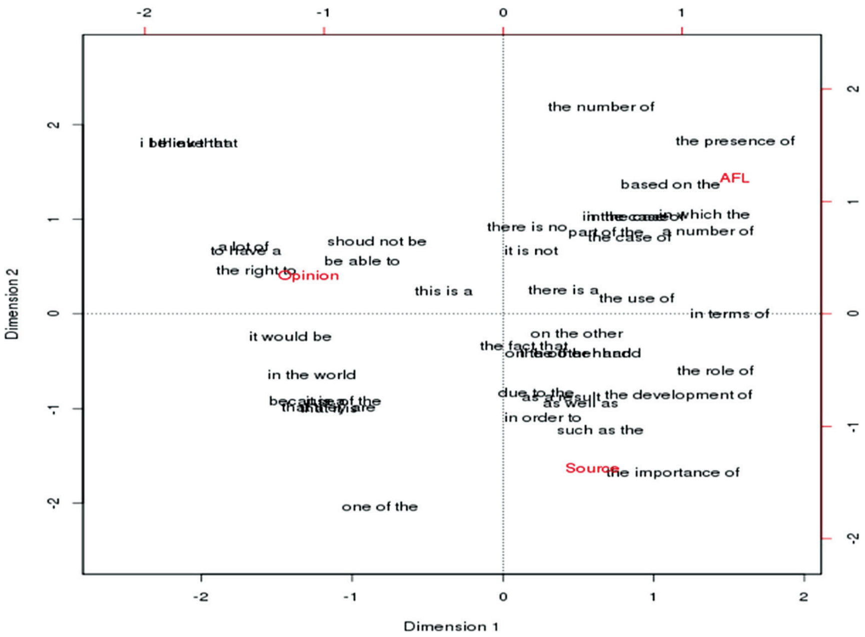


Figure 1. Correspondence analysis of the opinion, source, and AFL bundles

sion 1 (horizontal line) and 2 (vertical line) have the following eigen values and contribution rates: Dimension 1: eigen value 0.27, contribution rate 88.7%; Dimension 2: eigen value 0.03, contribution rate 11.3%. The column scores (locations on the biplot) for opinion, source, and AFL are as follows:

Opinion: Dimension 1 = −1.08, Dimension 2 = 0.33

Source: Dimension 1 = 0.50, Dimension 2 = −1.37

AFL: Dimension 1 = 1.30, Dimension 2 = 1.21.

Table 4 presents the noticeable bundles and their row scores of opinion, source, and AFL.

Table 4. Characteristic bundles and row scores of opinion, source, and AFL

Bundles	Dimension 1	Dimension 2
I believe that	−2.08	1.81
the presence of	1.54	1.81
the importance of	1.13	−1.69
in order to	0.27	−1.09
the fact that	0.14	−0.33

Figure 1 shows that dimension 2 (vertical line) separates the opinion-based bundles from the source-based and AFL bundles. One feature of opinion-based bundles is that they are characterized by assertive stance bundles e.g., *I believe that*, as shown in the upper left (dimension 1: −2.08; dimension 2: 1.81) in Figure 1. Because *I believe that* is never used in the AFL, some opinion-based stance bundles are too subjective for academic written English. Meanwhile, the lower right in Figure 1 demonstrates that the stance bundle *the importance of* (dimension 1: 1.13; dimension 2: −1.69) is frequent in source-based essays. This highlights an interesting difference between stance bundles in source-based and opinion-based essays; source-based essays take an objective stance with an inanimate subject *the importance*, while opinion-based essays display a subjective stance with a personal subject *I*. This could be due to the source differences the two argumentative essays base their arguments on; opinion: the writer's opinion or knowledge, source: objective evidence such as research articles. The similarity in frequency between *the importance of* with AFL (located on the right of dimension 1) suggests that the academic tone of source-based essays is closer to AFL. This aligns with Granger (2017), who found that noun-based bundles are a feature of academic writing. Another similarity of the source-based bundles with the AFL is the frequency of discourse bundles such as *in order to*. While this discourse bundle is located slightly on the lower right (dimension 2: −1.09), which shows the bundle's specificity to source-based essays, it has the potential applicability to academic written English, as it is also placed on the right of dimension 1. A moderate correlation (*r* = .60) of the top 21 source-based bundles with those in the AFL in frequency also reinforces their potential utility.

The upper right in Figure 1 implies that the AFL is distinguished by more objective noun-based bundles (e.g. *the number/presence of*) than opinion/source-based essays that have noticeable stance bundles such as *I believe that* and *the importance of*. This difference in argumentative tone should be considered in the applications of essays to academic writing.

Placed near the center in Figure 1 (dimension 1:0.14; dimension 2: −0.33), the referential bundle *the fact that* is commonly used regardless of text types (essays and research papers). This implies that *the fact that* is an objective and widely applicable academic bundle, which makes it an essential focus in the early stages of EGAP writing instruction.

In sum, the correspondence analysis revealed that (1) opinion-based bundles, especially stance (e.g., *I believe that*) ones, are too subjective and may not be suitable for academic written English ; (2) source-based bundles are more similar to academic written English than opinion-based bundles, as shown in the high frequency of discourse bundles such as *in order to*, *as well as*, and (3) referential (e.g., *the fact that*) in argumentative essays are widely applicable to academic written English.

To gain deeper insights into the bundle match rates and detailed frequency information, Table 5 provides the top 21 frequent bundles of the EBL (opinion/source) and the corresponding AFL bundles by frequency per million words.

The frequency information of each bundle in Table 5 strengthens the points discussed in the result of the correspondence analysis. First, the top two discourse bundles (*in order to*, *as well as*) in source-based essays bear a strong similarity with those in the AFL. Interestingly, the two bundles exhibit the same frequency order in the source and the AFL, with *in order to* being followed by *as well as*. The prominence of these two discourse bundles illustrates one feature of academic writing, which utilizes

Table 5. Top 21 lexical bundles in opinion, source, and the AFL

Opinion	Freq.	Source	Freq.	AFL	Freq.
be able to	567	**in order to**	487	in terms of	282
I think that	524	**as well as**	411	the use of	270
that it is	495	**the fact that**	298	**in order to**	255
a lot of	483	*one of the*	295	**as well as**	255
one of the	412	the use of	262	the number of	246
to have a	390	in terms of	260	**there is a**	223
in order to	388	**there is a**	233	part of the	216
the fact that	329	due to the	222	a number of	215
it would be	283	*that it is*	216	**the fact that**	203
it is a	276	as a result	207	**it is not**	188
as well as	269	on the other	180	**there is no**	185
it is not	252	such as the	178	the case of	168
there is no	248	**it is not**	173	in which the	166
there is a	243	part of the	168	in the case	153
I believe that	243	*be able to*	167	in the case of	135
the right to	243	the other hand	160	based on the	134
should not be	233	on the other hand	159	the presence of	130
that they are	217	the importance of	156	due to the	127
this is a	212	a number of	154	as a result	125
because of the	202	the development of	154	the development of	121
in the world	198	**there is no**	152	the role of	121

Italic = shared in opinion and source, shading = shared in source and AFL, **bold** = shared in all the three

the two bundles to create or organize logical connections of information. *In order to* also frequently appears in opinion-based essays, which suggests the bundle's adaptability to academic writing. Another similarity between source-based bundles and those in the AFL is the frequency of noun-based bundles. *In terms of* and *the use of* are ranked within the top 6 in both source-based essays and the AFL. This again demonstrates the high applicability of source-based bundles. Second, the referential bundle *the fact that* is shared in all the three (opinion, source, and AFL), meaning it is a common bundle applicable to a range of academic writing texts.

Dividing the EBL According to Difficulty Level

RQ 2 of the study examined the possible divisions of the EBL (3, 768 words) to facilitate its use in EGAP writing practice. As Nation (2016) pointed out, wordlists with numerous numbers of words (e.g., 1,000 words long) are too extensive to incorporate into a particular curriculum or course. Consequently, this study classified the EBL based on difficulty level of each bundle.

The CEFR and English Vocabulary Profile

To gather information on the bundles' difficulty level, this study referred to the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) and the English Vocabulary Profile (EVP; Capel, 2015). The CEFR categorizes foreign language learners' proficiency into six levels: beginner (A1), elementary (A2), intermediate (B1), upper-intermediate (B2), advanced (C1), and proficiency (C2), with A1 being the lowest and the C2 being the highest. The EVP utilizes actual learner-produced data (essays) to offer CEFR-based difficulty levels for phrases (lexical bundles). For example, the EVP states that A2 learners are expected to productively use the bundle *it is true that* in writing; thus, the bundle is at A2 level.

The three-to-five 3,768 bundles in the EBL were manually checked with the corresponding CEFR levels in the EVP. For appropriate difficulty levels, the classification was limited to A2 (elementary), B1 (intermediate), and B2 (upper-intermediate). This aligns with previous studies that focused on the Asian university students at these levels, including Japanese (Nam & Park, 2020; Sawaguchi, 2024). Table 6 shows the EBL divided into A2, B1, and B2 levels.

Table 6. CEFR-labelled EBL bundles

CEFR level	No. of bundles	Proportions (%)
A2	20	15.7%
B1	40	31.4%
B2	67	52.7%
Total	127	99.8%

Note: Percentages may not sum to exactly 100% due to rounding.

As shown in Table 6, B2 (upper-intermediate) level bundles occupy the largest proportion of the labelled CEFR levels. In the previous studies that targeted Asian university students (Nam & Park, 2020; Sawaguchi, 2024), B2 level students are considered the most proficient. This suggests that, overall, the EBL has challenging learning items for average Japanese university students. These include the bundles with relatively advanced vocabulary, including *the distinction between* whose content word *distinction* is at 5,000 level in the New Word Level Checker (NWLC; Mizumoto, 2021). The abundance of B2 bundles in source-based essays contributes to the overall large number of B2 bundles (52 of 67). In comparison, A2 (elementary) bundles account for the smallest proportion of the total labelled bundles (20 of 127). Some of them are characterized by basic vocabulary (e.g., *the fact that*, *it is true that*).

The proportional features of B1 and B2, which account for over 80% of the total A2, B1, and B2 bundles show the overall tendency that the EBL frequently employ the bundles that help university students present clear and logical arguments on various topics. This competency is in line with the CEFR descriptors of B1 "give reasons and explanations for opinions" and B2 "produce clear, detailed text on a wide range of subjects" (Council of Europe, 2001, p. 24), which again enhances the EBL's value to improve university students' basic academic writing skills.

Applying the EBL to the EGAP Writing Practice

Building on the CEFR categorization, the result of RQ 1 (discourse functions and the frequency of the EBL bundles and their similarities to the AFL), and the relevant findings in previous studies, I will discuss the applications of the EBL to EGAP essay writing activities. Model answer sentences were generated by ChatGPT 4o, and later modified by the author. Figure 2 illustrates how opinion-based bundles can be used in

The Debate on Free University Education: A Policy Worth Considering

Introduction: **Whether or not** free university education is a viable policy remains a contentious issue…. **it is true that** the financial cost of such a policy would be enormous. However, this **does not mean**...

Body 1: **To begin with**, free university education would lead to a more equitable society….**the fact that** some European countries, such as Germany and Sweden, have already implemented free university education….

Body 2: The amount of money needed to sustain free university education is another important consideration…

Conclusion: **All in all**, it is clear that such a policy would have positive effects….

Answer: A2: *it is true that*, *the fact that*
B1: *whether or not*, *does not mean*
B2: *to begin with*, *all in all*

Figure 2. Application of opinion-based bundles

teaching A2 and B1 university students. The answers with CEFR levels are also provided.

As discussed in RQ 1, *the fact that* is a commonly used academic referential bundle. It would be effective to focus first on the bundle. It is at A2 level in the EVP, whose literal meaning and lower level of vocabulary *fact* would facilitate A2 students' use of the bundles. The stance bundles *it is true that* and *does not mean* would also be useful to develop strong arguments in writing. As discovered in RQ 1, stance bundles with personal subjects (e.g., *I believe that*) are too subjective for academic writing; thus, using inanimate subjects *it* and *this* as in Figure 2 assists in maintaining objective academic tone. In fact, the *it is* construction is frequently used in the AFL (e.g., *it is important/necessary/possible* to…). Regarding discourse bundles (*whether or not*, *to begin with*, *all in all*), all of them are ranked at B1 or B2 in the EVP. These bundles can be considered appropriate difficulty for B1 and B2 learners. These discourse bundles can also be effective in academic writing; as RQ 1 found that discourse bundles in opinion-based essays (e.g., *in order to*) show a high frequency similar to that of AFL (ranked within the top 21).

The above suggestions for bundles in terms of difficulty and discourse functions are also supported by previous studies. Chen & Baker (2016) found that B2 students use more objective stance bundles with *it is* constructions, and Sawaguchi (2024) discovered the B1 students' competency development to employ varied discourse bundles (e.g., *it is up to*) compared to A2 students. Opinion-based stance bundles with inanimate subjects (e.g., *it is true that*, *does not mean*) are beneficial for A2 students to be aware of academic stance tone at the early stages of writing practice. B1 students can also increase their repertoire of discourse bundles with the focus on those at B1 and B2 levels (e.g., *whether or not*, *to begin with*, *all in all*).

Figure 3 presents the application of source-based bundles for B2 students.

It was found in RQ 1 that source-based essays frequently employ discourse bundles (e.g., *in order to*), which facilitate the organization of presenting information. For B2 students, continued focus on formal B2 discourse bundles like the ones in Figure 3 (*despite the fact that*, *one the one hand*) will assist B2 students in presenting their arguments more logically, because the two bundles contrast both sides of arguments in an objective manner. At B2 level, the effective use of advanced vocabulary is also necessary. As Figure 3 shows, the referential bundles with relatively advanced

> **Renewable Energy versus Fossil Fuels: A Choice for the Future**
>
> Introduction: As climate change impacts grow, the debate over energy sources is more urgent than ever. **Despite the fact that** fossil fuels have been central to global energy, renewable sources are increasingly viewed as essential for sustainability.
>
> Body 1: **On the one hand**, fossil fuels have been deeply embedded in global economies for centuries, but on the other,…
>
> Body 2: However, **the existence of** renewable energy in our current energy mix is still limited.
> …While fossil fuels and renewable energy are often compared on environmental grounds, **the distinction between** the two also lies in their economic implications….**The origins of** fossil fuels…
>
> Conclusion: In conclusion, while fossil fuels remain essential to the global energy supply, renewable energy can be a viable and sustainable energy source for the future.
>
> Answer: B2: *despite the fact that, one the one hand, the existence of*, *the distinction between*, *the origins of*

Note: References are required in actual source-based argumentative essays.

Figure 3. Application of source-based bundles

levels of content words (*the existence of*, *the distinction between*, *the origins of*) are at B2 level in the EVP. One feature of academic written English (AFL) is the frequent use of various referential bundles, including *the presence/development of*. Aiming at the referential bundles such as *the existence of*, *the distinction between*, and *the origins of* will further increase B2 students' use of sophisticated referential bundles.

Conclusion

This study sought to develop an initial framework for the list of lexical bundles for argumentative essay writing and to explore the list's potential applications to EGAP practice.

RQ 1 found that approximately 80 % of the lexical bundles in the EBL overlapped with those in the AFL, which suggests the potential for the application of argumentative essay writing to EAP. The analyses of the highly frequent top 21 bundles in the EBL and the AFL revealed the following: Opinion-based essays contain remarkable stance bundles such as *I believe that*, which may not be used in academic written English practice due to their subjectivity, while the objective referential bundles including *the fact that* is applicable. In contrast, source-based essays have the abundant discourse bundles (e.g., *in order to*, *due to the*), which are more similar to academic written English. Academic written English (AFL) is distinguished from both types of essays, with more frequent use of noun-based referential bundles (e.g., *the number/presence of)*.

RQ 2 classified the EBL into the CEFR-based difficulty (A2, B1, B2) level, suggesting appropriate bundles to teach at each proficiency level. Specifically, the basic referential bundles (e.g., *the fact that*) and objective stance bundles (e.g., *it is true that*) can be appropriate at A2 and B1 levels; discourse bundles (e.g., *to begin with*, *on the one hand*) should be the continued focus from B1 to B2 levels. Advanced referential bundles such as *the existence of* can be taught at B2 level.

Finally, the limitations of this study and the directions for future research are discussed. While the findings highlight the relevance of the EBL to EGAP instruction, further validation and adjustments are needed to refine the list and confirm its pedagogical effectiveness; thus, the list will not be publicly released at this stage. As this study represents the first attempt to develop a collection of essay-specific bundles, the study serves as a foundation for future research on the practicality of the EBL in EGAP, contributing to the development of argumentative and academic writing instruction.

Acknowledgements

This study was supported by JST SPRING (grant number JPMJSP2150).

References

Anthony, L. (2023). AntConc (Version 4.2.4) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Biber, D. (1989). A typology of English texts. *Linguistics, 27*(1), 3–43.

Biber, D., Conrad, S., & Cortes, V. (2004). “If you look at ...”: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*, 371–405.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*, 263–286.

Blue, G. (1988). Individualising academic writing tuition. In P. C. Robinson (Ed.), *Academic writing: Process and product* (ELT Documents 129, pp. 95–99). Modern English Publications.

Bychkovska, T., & Lee, J. J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes*, *30*, 38–52.

Capel, A. (2015). The English vocabulary profile. In Harrison, J., & Barker, F. (Eds), *English profile in practice*. Cambridge University Press.

Chen, Y. H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics*, *37*(6), 849–880.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*, 397–423.

Council of Europe. (2001). *The common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Crossley, S.A., Tian, Y., Boffour, P., Franklin, A., Benner, M., & Boser, U. (2024). A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Assessing Writing*, *61*, Article 100865.

Gardner, S., Nesi, H., & Biber, D. (2018). Discipline, level, genre: Integrating situational perspectives in a new MD analysis of university student writing. *Applied Linguistics*, *40*(4), 646–674.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (Ed.), *Learner English on computer* (pp. 3–18). Addison Wesley Longman.

Granger, S. (2017). Academic phraseology: A key ingredient in successful L2 academic literacy. In Vatvedt Fjeld, R., Hagen, K., Henriksen, B., Johannson, S. Olsen, S., & Prentice, J. (Eds.), *Academic Language in a Nordic Setting: Linguistic and Educational Perspectives*. *Oslo Studies in Language*, *9*(3), 9–27.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4–21.

Ishikawa, S. (2023). *The ICNALE guide: An introduction to a learner corpus study on Asian learners' L2 English*. Routledge.

Johnson, D. (2018). Teaching English for academic purposes in New Zealand: Making sense of genre-based instruction. In Wong, L.T., & Wong, Heidi. W. L (Eds.), *Teaching and learning English for academic purposes: Current research and practices* (pp. 239–253). Nova Science Publishers.

Mizumoto, A. (2015). Langtest. (Version1.0) [Web application]. https://langtest.jp/

Mizumoto, A. (2021). New Word Level Checker [Web application]. https://nwlc.pythonanywhere.com.

Nam, D., & Park, K. (2020). Lexical bundles as criterial features in L2 academic writing: structural differences between CEFR A2 and B2 essays. *Multimedia-Assisted Language Learning*, *23*(3), 68–86.

Nation, I.S.P. (2016). *Making and using wordlists for language learning and testing*. John Benjamins Publishing company.

Nesi, H., Gardner, S., Thompson, P., & Wickens, P. (2008). British academic written English corpus (BAWE). Centre for Corpus Research, University of Birmingham.

Nesi, H., & Matheson, N., & Basturkmen, H. (2017). University literature essays in the UK, New Zealand and the USA: Implications for EAP. *New Zealand Studies in Applied Linguistics*, *23*(2), 25–38.

Omidian, T., Shafriari, H., & Siyanova-Chanturia, A. (2018). A cross-disciplinary investigation of multi-word expressions in the moves of research article abstracts. *Journal of English for Academic Purposes*, *36*, 1–14.

Römer, U., & O'Donnell, M. B. (2011). Michigan corpus of upper-level student papers (MICUSP). University of Michigan, Ann Arbor.

Sawaguchi, R. (2024). Potential of L1 and L2 corpora to identify target lexical bundles for argumentative essay writing. *Asia Pacific Journal of Corpus Research*, *5*(1), 1–21.

Shoufaki, S., & Petrić, B. (2021). Academic vocabulary in an EAP course: Opportunities for incidental learning from printed teaching materials developed in-house. *English for Specific Purposes*, *63*, 71–85.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*(12), 487–512.

Swales, J.M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Tardy, C. M., Buck, R., Jacobson, B., LaMance, R., Pawlowski, M., Slinkard, J. R., & Vogel, S. M. (2022). "It's complicated and nuanced": Teaching genre awareness in English for general academic purposes. *Journal of English for Academic Purposes, 57*, 101117.

Wingate, U. (2012). 'Argument!' helping students understand what essay writing is about. *Journal of English for Academic Purposes*, *11*(2), 145–154.

Wu, S. M. (2006). Creating a contrastive rhetorical stance: Investigating the strategy of problematization in students' argumentation. *REC journal*, *37*(3), 329–353.

Yoon, H.J., & Tabari, M.A. (2023). Authorial voice in source-based and opinion-based argumentative writing: Patterns of voice across task types and proficiency levels. *Journal of English for Academic Purposes*, *62*, Article 101228.

（Ryo Sawaguchi, Graduate School of Foreign Language Education and Research,
Kansai University: rswgch@gmail.com）

「論文」

Work Your Way through Authentic Data: Data-driven Construction Learning and Its Effectiveness Explored through an Experimental Study

Daisuke MANABE

Abstract

Corpora have been utilized for purposes of language pedagogy. One of the approaches, data-driven learning (DDL), uses corpora or corpus-based materials in language classrooms. DDL is an inductive language learning method in which learners explore authentic language data and discover linguistic patterns on their own. While empirical studies on DDL have been increasing (Boulton & Cobb, 2017), there are only a few experiments on “data-driven construction learning” (Gilquin, 2021). The present paper reports the results of an experiment aimed at testing the effectiveness of data-driven construction learning and evaluating learners’ attitudes towards DDL. In the experiment, two groups of Japanese learners of English learned the way construction (e.g., Goldberg, 1995; Luzondo Oyón, 2013) with one group learning through DDL and the other through a traditional form-focused instruction. DDL in this study includes an explicit explanation of the target construction, following the tenet of applied construction grammar (Gilquin & De Knop, 2016). The effectiveness of the two methods was measured and compared by means of pre- and post-tests (sentence production and translation tasks). Additionally, the learners’ attitudes towards DDL were investigated through a post-questionnaire. The improvement of both sentence production and translation tasks in the post-tests demonstrated that both DDL and the traditional instruction were effective. Also, the participants’ attitude towards DDL was found to be positive. However, learners who received the traditional instruction outperformed those who received DDL. Therefore, the present study concludes that even though DDL was effective, other teaching methods could be more beneficial for learners, depending on the difficulty of a target construction and learners’ proficiency. This paper also argues that learners can benefit from DDL in various ways, such as developing general cogni-

tive skills, and hence it is suggested that incorporating DDL into a classroom activity or employing it as an out-of-class activity would be advantageous.

1. Introduction

Corpora have been extensively utilized in language pedagogy (cf. Leńko-Szymańska & Boulton, 2015). One of the approaches for foreign language learning based on corpora is data-driven learning (DDL). DDL uses corpora to facilitate foreign language learning, and this method allows learners to interact with authentic language data and discover linguistic patterns on their own (i.e., inductive language learning), using corpora or corpus-based materials. DDL can be categorized into two types according to the way corpora are used. The first type is computer-based or direct DDL in which learners have direct access to corpora. The second type is paper-based or indirect DDL in which learners use corpus-based materials and they do not have access to a corpus (see Yoon & Jo, 2014; Gilquin & Granger, 2022 for the clear distinction between direct and indirect use of corpora in DDL). Through DDL, learners can receive considerable amount of linguistic input by being exposed to a large number of authentic instances of a target lexical or grammatical item. DDL not only helps learners to become aware of linguistic patterns in their second language (L2) but also develops general cognitive skills for language learning (O' Sullivan, 2007, p. 277; Yoon & Jo, 2014, pp. 96-97). DDL has gained substantial attention and there have been a number of DDL studies since the approach was introduced by Johns (1991), one of the innovators of DDL. Meta-analyses have demonstrated that the effectiveness of the method was evident across multiple studies (Boulton & Cobb, 2017; Mizumoto & Chujo, 2015). The main focus of the application of DDL, however, has been on learning lexical and lexico-grammatical items, and "(larger) units such as constructions, by contrast, tend to be neglected" (Gilquin, 2021, p. 230) in DDL studies. Learning constructions (a construction is a conventionalized pairing of form and meaning as defined in construction grammar; e.g., Goldberg, 2006) through DDL is called "data-driven construction learning" (Gilquin, 2021). Specifically, to the best of my knowledge, there are no studies on data-driven construction learning targeting Japanese learners of English, other than Manabe (2024). Since empirical studies on data-driven construction learning have been rarely conducted, the present paper will focus on applying DDL to the learning of an abstract syntactic pattern by Japanese learners of English from a constructionist perspective. The aim of the present paper is to test the effectiveness of DDL in construction learning for Japanese learners of English by means of pre- and post-tests, and to also investigate learners' attitudes towards DDL via a post-questionnaire. In this experimental study, two groups of participants learned the *way* construction (e.g., Frank dug his way out of the prison; Goldberg, 1995, p. 199; see also Luzondo Oyón, 2013). Learning the *way* construction can benefit learners of English, as it facilitates both reading comprehension and natural expression of progress or movement in communication[1]. One of them learned the target construction through DDL, and the other learned through traditional form-focused instruction. The three main research questions are addressed in the current study:

1. Can Japanese learners of English effectively learn the *way* construction through DDL?
2. Is DDL equally or more effective in construction learning compared to a traditional form-focused instruction?
3. Do Japanese learners of English show positive attitudes towards DDL?

The structure of this paper is as follows. Section 2 introduces the theoretical background, i.e., construction grammar, and "data-driven construction learning" (Gilquin, 2021, p. 231) which applies DDL from the constructionist perspective. In Section 3, the experiment conducted at a national university in the Chubu region of Japan and the data analysis methods are described. In Section 4, the results of the pre- and post-tests and the evaluation of DDL are presented. Finally, Section 5 argues that, despite the fact that the traditional instruction group outperformed the DDL group, DDL has great potential for improving linguistic knowledge and other cognitive skills.

2. DDL and Construction Learning

2.1 Construction Grammar

A construction in construction grammar (e.g., Goldberg, 2006; Hilpert, 2019; Hoffmann, 2022; Hoffmann & Trousdale, 2013) is a basic linguistic unit that has a form-meaning pair. Any level of linguistic item (e.g., morphemes, words, idioms, argument structure constructions) is seen as a construction if they contain a conventionalized pairing of form and function (Goldberg, 2006, p. 3). Linguistic knowledge in speakers' minds forms a large network of constructions (Hilpert, 2019, p. 2). A usage-based framework claims that the constructional network is built through generalizing a huge amount of linguistic input (Gilquin, 2021, p.231). Usage-based theories suggest that language is acquired through actual language use, and therefore linguistic input and frequency are regarded as a crucial factor for language acquisition (Hoffmann, 2022, p. 27). From a usage-based perspective, DDL that can provide a considerable amount of authentic input is expected to be effective.

In applied construction grammar (Gilquin & De Knop, 2016), it is considered what speakers learn when acquiring a foreign language is constructions (Gilquin, 2016, p. 146). Thus, L2 learners acquire constructions of a target language during L2 acquisition. The acquisition of a first language (L1) and L2 is different in several ways, and the differences are attributable to learning environments, amount of input, authenticity of input, learning process (inductive/implicit vs. deductive/explicit), and so forth (Gilquin,2021, pp. 231-232). However, by adopting DDL, learners can be exposed to a substantial amount of authentic input and inductively learn target constructions. Consequently, it is possible that DDL brings the process of L2 learning closer to that of L1 acquisition (Gilquin, 2021, p.231).

2.2 Data-driven construction learning

Gilquin (2021) applied DDL to learning constructions from the perspective of usage-based construction grammar, and called this approach "data-driven construction learning" (p. 231). In data-driven construction learning, the focus is primarily on an abstract syntactic pattern. In the experiments of Gilquin (2021), high-intermediate learners of English studied three constructions (i.e., the MAKE causative construction, the *way* construction, and the *into* causative construction). As a result of pre- and post-tests, the participants demonstrated a strong understanding of the target constructions. After DDL, an increase in the number of produced sentences and an improvement in the quality of the sentences (native-like quality) were observed. Additionally, the use of "new verbs" (Gilquin, 2021, p 238), whose cooccurrence with the target constructions was not introduced in the DDL material, was found in the produced sentences of the *way* construction and the *into* causative construction. This suggests that DDL led learners to generalization of knowledge of the two constructions. The same phenomenon was also observed in the learning of the *way* construction by Japanese learners of

English in Manabe's (2024, p. 23) experiment. However, Gilquin (2021) also pointed out some downsides of DDL such as the lack of long-lasting effect and the time-consuming nature of the method (p. 242). Similarly, several previous studies have highlighted both the same and other weaknesses of DDL (e.g., Chambers, 2022, p. 420; Boulton, 2010, pp. 535-537; Gilquin & Granger, 2022, p.436). As for the evaluation of DDL, it was reported that both positive and negative attitudes towards DDL were observed and it was pointed out that DDL was not favored by some learners (Gilquin, 2021, p. 241). Manabe (2024, p. 24) reported Japanese learners' positive attitudes towards construction learning through DDL. Furthermore, learners' positive attitudes towards DDL were evident in a number of previous studies (e.g., Boulton, 2010, p. 557; Gilquin & Granger, 2022, p.436; Mizumoto & Chujo, 2015, p. 12; Takahashi & Fujiwara, 2016, p. 95).

3. The Experiment

3.1 Experimental design

The experiment was composed of pre- and post-tests, a pre- and post-questionnaire, and the educational intervention. The experiment started with the pre-questionnaire (about five minutes) followed by the pre-tests (16 minutes). After a 10-minute break, the participants received the educational intervention for 30 minutes, which was followed by another 10-minute break. Then the participants took the post-tests (10 minutes) and completed the post-questionnaire (no time limit).

3.2 Participants

The participants were L1 Japanese speakers at a national university in the Chubu region of Japan, who were learning English as a foreign language. Forty students (37 undergraduates and three graduates) took part in the experiment (M age = 20.4 years, SD = 1.96; M years of English language learning experience = 9.8 years, SD = 3.15). Due to the random sampling procedure, participants' proficiency levels varied considerably, ranging from A2 to C1 on the CEFR scale, with B1 being the most frequent level[2]. The participants were divided into two groups: the DDL group and the traditional instruction (TI) group.

3.3 Educational interventions

The participants learned the *way* construction (e.g., Goldberg, 1995; Luzondo Oyón, 2013) using a concordance and a worksheet (henceforth, the DDL material) created by the author (see Appendix 1 and Appendix 2). The concordance consisted of twenty instances of the *way* construction extracted from the Corpus of Contemporary American English (COCA: Davis, 2008-). The author selected sentences that seemed relatively easy for learners to understand, based on vocabulary (e.g., the absence or presence of technical terms) and the length of the sentences.

The DDL group received a paper-based DDL. DDL in this study includes an explicit explanation of the central form and meaning of the target construction. I will refer to this DDL approach as construction-centered DDL. A pilot study[3] and a previous study (Manabe, 2024) showed that the *way* construction is a difficult construction for Japanese learners of English. Hence, construction-centered DDL was developed because it was expected that an explicit explanation of the construction would facilitate learners' understanding of the *way* construction (see Sung & Yang, 2016 for the effects of construction-centered instruction). In the DDL intervention, the author first briefly explained DDL and how to interpret the concordance prior to students' independent learning of the target construction. The participants were asked to read example sentences in the concordance and work on the worksheet. The tasks on the worksheet included translating into Japanese, paraphrasing, and describing forms and meanings that learners discovered (see Appendix 2). To eliminate the possibility that other factors would influence learning outcomes, there was neither teacher intervention nor interaction among the participants. The participants were permitted to use a dictionary to look up words within the concordance. However, searching for the *way* construction was prohibited. After the DDL intervention, the DDL material was collected.

In the TI group, the participants learned the *way* construction in a more traditional way. The instruction was a teacher-centered lecture, mainly focusing on the form of the target construction. In the first task, namely a syntactic task, the participants categorized six sentences, which have the term "way," into three groups based on their forms. Then they were provided with an explanation of the *way* construction with Japanese translations and a few examples. After going through the explanation of the *way* construction, the participants completed three types of exercises: True or False, Sentence Scramble, Fill-in-the-blank. Finally, they had some time (up to five minutes) to individually review what they learned. The sentences in the TI material were also based on corpus data in order to ensure that participants would not receive any inappropriate input. The overall contents of each instruction are shown in Table 1.

Table 1. The overall contents of each instruction

DDL (30 minutes)	**TI (30 minutes)**
Introduction (An explanation of DDL and the concordance)	Introduction (Syntactic task)
DDL (The concordance with an explicit explanation about the *way* construction, and the worksheet)	Form-centered explanation (Japanese translations and examples)
	Exercise (True or False, Sentence Scramble, Fill-in-the-blank)
	Review
Collection of the material	Collection of the material

3.4 Pre- and post-tests

The pre-tests consisted of three types of tests: a vocabulary size test (VST; Hamada et al., 2021)[4], a sentence production task (SPT), and a translation task (TT). This study utilized the VST to examine participants' prior knowledge of English. In the SPT, the participants were asked to generate as many sentences as possible containing the *way* construction within five minutes. Since the *way* construction was considered a highly difficult construction and the term "the *way* construction" is not well-known, it is assumed that producing sentences using this construction had become unnecessarily difficult (Manabe, 2024, p. 22). Therefore, the form of the *way* construction (subject + verb + one's way + preposition/adverb) and two example sentences (i.e., "He made his way through the crowd" and "The kid crawled his way into the room") were provided in the SPT in the pre-tests. In the post-tests, the form and the example sentences of the *way* construction were removed. The TT was conducted to investigate whether participants understood the meaning of the target construction. In the TT, the participants translated five English sentences about the *way* construction into Japanese within five minutes. The questions were generated by the author based on corpus data (see

Appendix 3). The TT in the pre- and post-tests are essentially identical, differing only in a few modified elements (e.g., subjects and possessive pronoun). For example, one paired question about "make one's way through" in the TT was "She made her way through the forest" in the pre-tests and "He made his way through the crowd" in the post-tests.

3.5 Pre- and post-questionnaires

The pre-questionnaire collected information about speaker attributes (e.g., age and proficiency). In the post-questionnaire, a Likert scale (5-point) and open-ended questions were included to investigate learners' attitudes towards DDL.

3.6 Analysis

The results of the SPT were analyzed based on correct and incorrect usage of the *way* construction. The correct and incorrect usage discussed in this paper evaluated the *way* construction, and other errors were not taken into consideration (e.g., errors in inflections). The analysis of the produced sentences was carried out in the following steps: (1) the verification of form and meaning, (2) a corpus-based confirmation, and (3) an appropriateness judgment by L1 English speakers. In Step 1, the form of the produced sentences was checked, and sentences that did not follow the "verb + one's way + preposition/adverb" structure were classified as incorrect usage. Sentences that conformed to the form of the *way* construction but did not have the semantics of the *way* construction were also classified as incorrect usage (e.g., "I will go my way to achieve my goal"). As the next step, the sentences remaining from Step 1 were searched in COCA. If an expression was found in COCA, the sentence was classified as correct usage. Finally, the sentences remaining from Step 2, totaling 119 sentences, were judged by four L1 English speakers[5]. The appropriateness was evaluated using a 4-point Likert scale, and sentences that received an average rating of 3 or higher were classified as correct usage. The answers of the TT were evaluated by two L1 Japanese speakers[6] (including the author), and only the answers that received consistent evaluations from both raters were classified as correct answers.

4. Results

4.1 Tests

The sentences produced in the pre- and post-tests, with a total of 298 sentences, were analyzed. All the statistical analyses were performed using R (R Core Team, 2024). The number of correct and incorrect sentences is shown in Table 2, and the proportion of them is illustrated in Figures 1 and 2.

The Shapiro-Wilk test was utilized to test the normality of the data, and its results are shown in Table 3. The results showed that only the TI's SPT score in the post-test followed a normal distribution. Since all the other scores showed non-normal distributions, non-parametric tests were deemed appropriate for the statistical analyses in the present study.

Table 2. The number of correct and incorrect sentences in the SPT in the pre/post-tests

	DDL group		TI group	
	correct	incorrect	correct	incorrect
Pre-test	4	42	6	37
Post-test	50	38	103	18

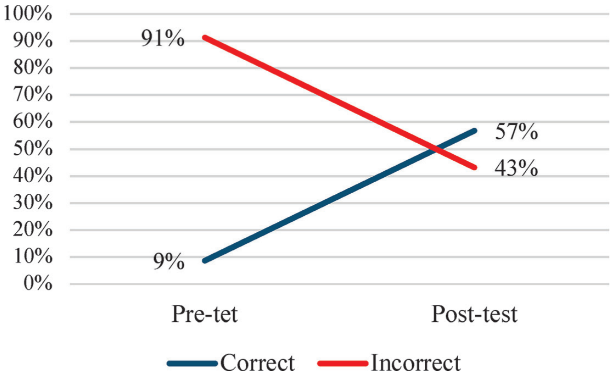


Figure 1. The proportion of correct and incorrect sentences in the SPT of the DDL group

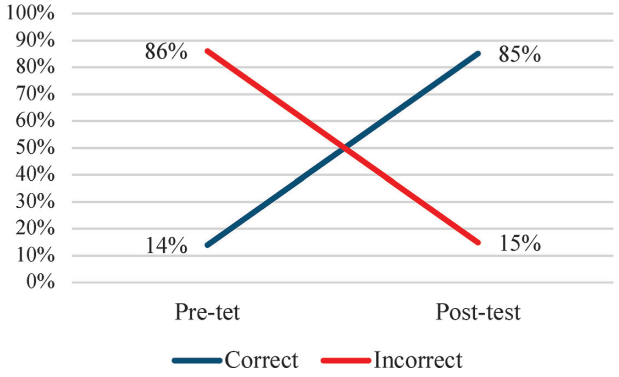


Figure 2. The proportion of correct and incorrect sentences in the SPT of the TI group

First of all, it is worth noting that the participants in the two groups were at almost the same levels in terms of vocabulary and knowledge of the target construction. No significant difference was found between the DDL group and the TI group in the VST by the Mann-Whitney U-test (W = 209.5, p=0.8012, effect size r = 0.0406). Therefore, the participants in the two groups were likely at almost the equal levels of English proficiency in terms of vocabulary. Also, the Mann-Whitney U-test indicated that there was no significant difference between the DDL group and the TI group in both the pre-test SPT (W =180, p = 0.4820, effect size r = 0.0855) and the pre-test TT (W = 216, p = 0.6546, effect size r = 0.0684). Hence, the participants in the two groups

Table 3. The results of the Shapiro-Wilk tests

		W		p-value	
		DDL	TI	DDL	TI
Pre-test	VST	0.8526	0.8554	0.0059	0.0066
	SPT	0.4954	0.5804	< 0.001	< 0.001
	TT	0.7869	0.8100	< 0.001	0.0012
Post-test	SPT	0.9011	0.9572	0.0433	0.4902
	TT	0.8620	0.8496	0.0085	0.0052

appeared to have almost the same prior knowledge about the *way* construction.

According to the results of the pre- and post-tests, DDL was effective in learning the *way* construction. As for the SPT of the DDL group, the results of the Wilcoxon signed-rank sum test indicated a significant difference between the pre- and post-test (V = 3, p < 0.001, effect size r = 0.8515). The results of the TT of the DDL group also showed a significant difference between the pre-test and post-test by the Wilcoxon signed-rank sum test (V = 0, p = 0.0013, effect size r = 0.8765). In addition, as shown in Table 4, the eighteen participants (90% of the total) in the DDL group produced more correct sentences in the post-test than in the pre-test. However, there were two participants who did not improve in producing the target construction after DDL. These results suggest that the DDL intervention was effective to some extent, and almost all of the participants effectively learned the *way* construction through DDL.

Regarding the TI, the effectiveness of the instruction was confirmed. The results of the SPT of the TI group indicated a significant difference between the pre- and post-test by the Wilcoxon signed-rank sum test (V = 0, p < .001, effect size r = 0.8763). The results of the TT also showed a significant difference between the pre- and post-test by the Wilcoxon signed-rank sum test (V = 0, p < .001, effect size r = 0.8765). Additionally, all the participants in the group showed an increase in the number of correct sentences in the post-test SPT (see Table 5). As the results illustrate, the traditional form-centered lecture was also effective for learning the *way* construction.

To see whether there was a difference in effectiveness of the two teaching methods, the results of the post-tests (both the SPT and TT) were compared with the Mann-Whitney U-test. The results of the SPT showed that there was a significant difference between the DDL group and the TI group (W = 64, p < 0.001, effect size r = 0.5817).

Table 4. The number of correct sentences in the pre/post-test for each participant in the DDL group

	Participant ID																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Pre	1	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Post	5	5	5	2	5	2	2	2	1	0	1	0	3	3	2	2	4	1	1	4

Pre = the pre-test, Post = the post-test

Table 5. The number of correct sentences in the pre/post-test for each participant in the TI group

	Participant ID																			
	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Pre	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0	1	1	0
Post	4	6	4	8	3	5	3	9	1	4	6	4	4	4	6	6	7	5	8	6

Pre = the pre-test, Post = the post-test

Table 6. The difference of the proportion of correct answers in the TT in the post-tests

Expressions	DDL	TI	Difference (TI - DDL)
make one's way through	65%	95%	30%
find one's way to	20%	100%	80%
elbow one's way through	65%	70%	5%
talk one's way out of	15%	55%	40%
work one's way through	30%	85%	55%

The results of the TT also indicated a significant difference between the two groups (W =38, p < 0.001, effect size r = 0.6929). Furthermore, an analysis of the accuracy rates for the two groups in the post-test TT indicated that the TI group outperformed the DDL group on all five questions (see Table 6). Focusing on the accuracy rates of each question in the pre- and post-tests, an increase in accuracy rates for all five questions was observed in both the DDL group and the TI group (see Figures 3 and 4). For all five questions, the improvement in accuracy rates (subtracting the pre-test scores from the post-test scores) was greater in the TI group than in the DDL group. For example, the increase in the accuracy rate for “make one’s way through” is 15 % in the DDL group but 75% in the TI group. These results suggest that the TI was more effective than DDL in this experiment, and the participants in the TI group were able to learn the *way* construction more effectively.

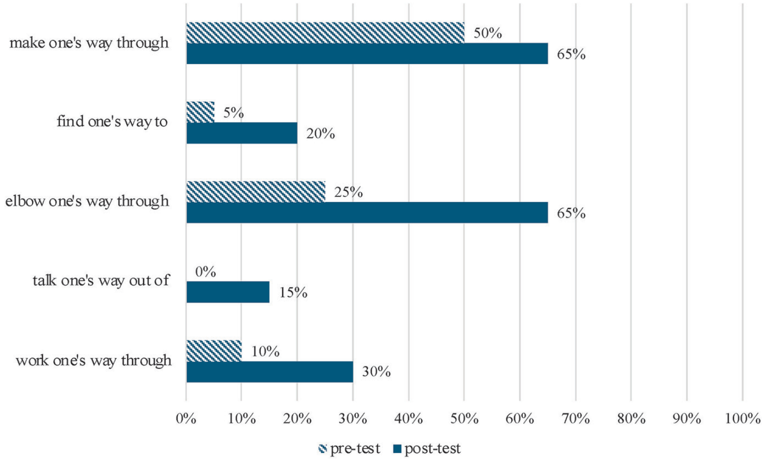


Figure 3. The accuracy rates for each expression in the TT in the pre/post-tests in the DDL group

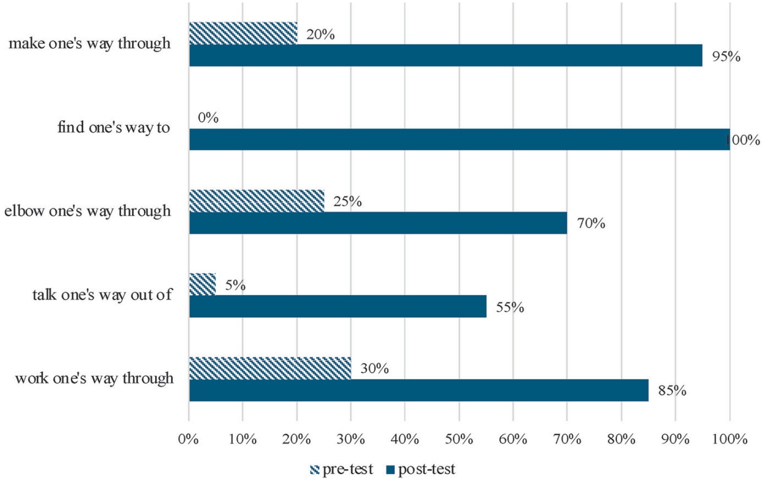


Figure 4. The accuracy rates for each expression in the TT in the pre/post-tests in the TI group

4.2 Evaluation of DDL

The post-questionnaire investigated how learners who experienced DDL perceived the approach. Figure 5 presents the questions measured on a 5-point Likert scale[7]. All the questions for which the sum of “agree” and “strongly agree” is 75% or more are related to positive opinions. There are some participants who felt that DDL and the tasks were difficult. However, there were hardly any negative opinions towards DDL. The participants tended to have positive attitudes towards DDL in the present study.

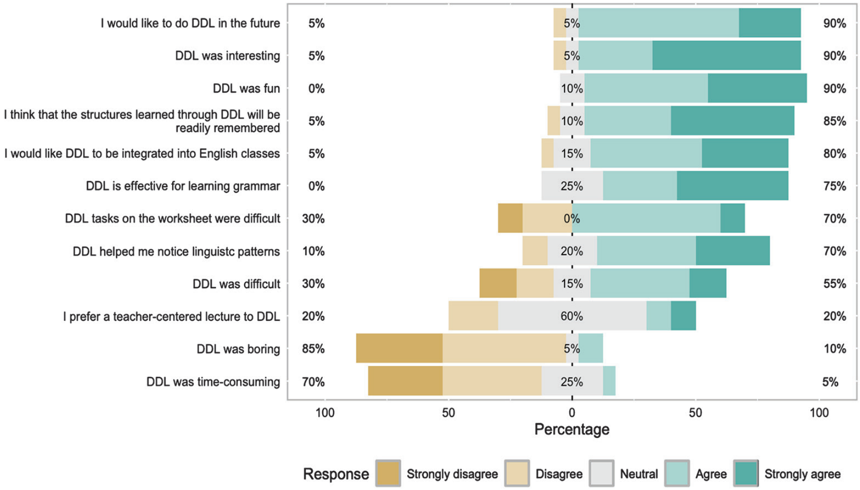


Figure 5. The results of the post-questionnaire in descending order based on the sum of “agree” and “strongly agree” (N=20)

In the open-ended questions, the participants were asked to describe the “positive aspects of DDL” and “aspects they disliked about DDL.” In the descriptions of the positive aspects, the participants mentioned the amount of examples, proactive learning, discovery, and so forth (Examples 1 and 2). In the responses regarding the aspects that participants did not like, worries about their understanding of the target construction (Examples 3 and 4) and the the lack of teacher intervention (Example 5), and so forth, were identified.

(1) The amount of example sentences was enough to understand the grammar.

(2) Since I discover the features myself, I can learn proactively.

(3) I was concerned that if my understanding was wrong, I might have learned it incorrectly.

(4) I get anxious about whether the similarities I discovered are actually correct.

(5) Since there was no explanation from the teacher, I didn’t know if what I was thinking was accurate.

5. Discussion

The present study illustrated that construction-centered DDL was effective in learning an abstract construction. The effectiveness of the method was measured by a comparison between the pre- and post-tests. As the analysis in Section 4.1 revealed, both the DDL group and the TI group showed an increase in correct sentences and in accuracy rates in the post-tests. The results indicate that DDL and the TI are effective, and that learners are able to capture the central form and meaning of the target construction through both methods. It can be concluded that construction-centered DDL has a positive impact on the learning of a construction. Thus the first research question, regarding the effectiveness of DDL, is framed positively.

While DDL conducted in this study was found to be effective, the TI was more effective in helping the participants learn the *way* construction. The comparison between the two groups showed that the TI group demonstrated a greater improvement in both SPT and TT in the post-test than the DDL group. While 18 participants (90% of the total) in the DDL group produced more correct sentences in the post-test than in the pre-test, all the participants in the TI group showed an increase in correct sentences. Also, the TI group illustrated higher scores in all five questions of the post-test TT than the DDL group. However, given the fact that Japanese translations were provided during the TI but not in DDL, this is not surprising. As the aforementioned results suggest, learners who receive the TI for learning the *way* construction are presumed to enhance their understanding and production of the construction more effectively than those who learn it through DDL. Hence, the analysis points to a negative response to the second research question regarding a comparison between DDL and the TI, because even though DDL was effective, the TI group outperformed the DDL group in both the SPT and TT.

One possible reason the TI group outperformed the DDL group may be attributable to the difficulty of the target construction. In total, there were only 10 (out of 89) correct sentences produced by the participants in the SPT in the pre-tests. Three sentences (about 33%) of the correct sentences in the pre-tests were imitations of the example sentences provided in the pre-tests, with only minor modifications, such as changes in the subject and possessive pronoun. For example, one of the produced sentences was "She made her way through the crowd," whereas one of the example sentences was "He made his way through the crowd". Only 10 out of 40 participants were able to produce a correct sentence with the *way* construction in the pre-tests. In addition, the TT in the pre-tests also indicated weak performance. The average accuracy rate of the TT for the DDL group was 18%, while that for the TI group was 16% (both groups answering five questions with 20 participants each). Taking these results into account, it is concluded that the *way* construction is a highly difficult construction for Japanese learners of English (possibly for learners of English with different L1s as well). Since the *way* construction is a difficult construction for Japanese learners of English, the participants in the DDL group might have struggled to understand the construction, produce them in their own words, and generalize what they learned through the input in the DDL material. This can explain why the two participants in the DDL group did not improve in sentence production (see Section 4.1). If this is the case, then there is a great possibility that the difficulty of a target construction will have a great impact on the effectiveness of DDL. Another explanation for the somewhat unfavorable results of the DDL group is that the explicit explanation of the target construction might have confused the participants. There might be a need to refine the explicit explanation, minimizing linguistic terms and making it comprehensible for any learners. To determine whether these are true, further empirical studies are required considering different levels of constructions as a learning target in DDL.

In addition to the difficulty of a target construction, learners' proficiency level should be taken into consideration. The present study did not include proficiency level in the analysis. Future studies should include learners' proficiency because it can also be a strong factor that affects outcomes of DDL. For advanced learners, simple exposure to input may be sufficient to learn a construction, as they are likely to have a sensitivity to discerning linguistic patterns. That is to say, advanced learners are able to extract patterns and generalize them by themselves. Lower-level learners, on the other hand, may not be sensitive enough to discern linguistic patterns, and hence have difficulties learning a target construction without assistance such as an explicit instruction.

As for the evaluation of DDL, it was found that the participants tended to have positive attitudes towards DDL, as the results of the post-questionnaire illustrated. Hence, the answer to the third research question is affirmative. Learners' positive attitudes towards a teaching/learning method are likely to contribute to positive outcomes (Gilquin, 2021, p. 239). As some participants mentioned (see Section 4.2), the considerable amount of instances were favored, and this is probably because grammar instruction usually includes fewer example sentences for the learning items. The more advanced learners are, the more likely they are to prefer autonomous learning, i.e., DDL, as they are capable of processing a large amount of input on their own. Some participants found DDL to be challenging, and this could be because they were not used to receiving a considerable amount of input in a short period. Also, some of the participants felt anxious while doing DDL because they did not know the answers to the questions in the worksheet or they sometimes did not comprehend some of the instances in the concordance. DDL as a classroom activity can involve teacher intervention and interaction among students, which will scaffold students' understanding of a target construction and reduce their anxiety. Accordingly, DDL may be more positively evaluated by a larger number of students.

Even though the DDL group did not perform as well as the TI group in the present study, the DDL group succeeded in capturing the target construction to some extent and the effectiveness of DDL was confirmed. Furthermore, it has been claimed that DDL can promote generalization of constructional knowledge that learners learn through DDL (Gilquin, 2021; Manabe, 2024), as shown in Section 2.2. This suggests that DDL is effective not only for rote memorization but also for the generalization of linguistic knowledge. In addition, previous studies claimed that DDL can develop general cognitive abilities, including "predicting, observing, noticing, thinking, reasoning, analysing, interpreting, reflecting, exploring, making inferences (inductively or deductively), focusing, guessing, comparing, differentiating, theorising, hypothesising, and verifying" as listed by O' Sullivan (2007, p. 277). These skills "may also be transferred to other fields of study" (Gilquin & Granger, 2022, p. 431). Another useful application of DDL is error correction (Gilquin & Granger, 2022, p. 430). Through DDL in a classroom, learners can become familiar with corpus consultation and eventually they will be able to autonomously utilize corpora whenever needed, such as for academic writing. The implementation of DDL is beneficial for learners in terms of developing such skills. DDL is therefore a method with great potential to help learners of a foreign language advance in terms of linguistic knowledge and various other skills. Moreover, the effectiveness of DDL can be fostered by combining it with other teaching methods (Gilquin, 2021, p. 243). Incorporating DDL into classroom activities can help learners acquire a target construction and get used to its authentic usage. Also, DDL can be carried out outside of the classroom. For example, learners can work on DDL materials as homework, and then discuss their discoveries in pairs or groups in the classroom. Teachers can scaffold their understanding by asking questions and having them complete extra projects during the class. DDL is originally designed for autonomous learning, having learners independently explore linguistic data. Hence, it is also suitable as an out-of-class activity. This may make it easier for teachers to adopt DDL in their classes.

6. Conclusion

This study demonstrated the effectiveness and potential of data-driven construction learning as well as possible refinements to the method. Several limitations of this study were also pointed out, such as the excessive difficulty of the target construction, the exclusion of proficiency levels in the analysis, and heterogeneity in proficiency levels (see Section 3.2). In future empirical studies, proficiency levels and different levels of constructions as learning targets must be taken into account to determine whether these factors have a significant impact on the effectiveness of DDL. Additionally, ways to incorporate DDL into an actual classroom must be explored. For example, integrating generative AI into DDL is one possible future direction (see Crosthwaite & Baisa, 2023; Mizumoto, 2023 for the synergy between AI and DDL). I hope this study will contribute to future research on DDL and its dissemination in educational settings.

Footnotes

1. Iida (2021, p. 112) pointed out that the *way* construction warrants pedagogical attention, as it appears in high school English, such as in university entrance examinations and some textbooks.
2. The participants' English proficiency levels were determined based on English

proficiency tests (e.g., TOEIC and IELTS), which were converted to CEFR (Ministry of Education, Culture, Sports, Science and Technology, 2015). The distribution was as follows: 1 participant at A2 (2.5%), 19 participants at B1 (47.5%), 10 participants at B2 (25%), and 2 participants at C1 (5%). Among the total participants, proficiency data were unavailable for 8 participants (20%) who had not submitted their test scores.

3. The pilot study was conducted with six participants (three undergraduates and three graduates) in January and February 2024. Only one participant was able to produce correct sentences with the *way* construction in the pre-test sentence production task.
4. Levels 1 to 3 (60 items) of the VST (Hamada et al., 2021) were used.
5. The four L1 English speakers were three Americans and one Australian.
6. Two L1 Japanese speakers (including the author), who have knowledge of the *way* construction, evaluated the answers of the translation task.
7. The original questionnaire and responses to open-ended questions by the participants were in Japanese but they were translated into English by the author without changing the meaning.

Acknowledgement

My sincere thanks go to all the participants and informants who made this research possible. This work was supported by JST SPRING, Grant Number JPMJSP 2135.

References

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

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-Analysis. *Language Learning*, *67*(2), 348–393.

Chambers, A. (2022). What is data-driven learning? In O'Keeffe, A. & McCarthy, M. (Eds.) *The Routledge handbook of corpus linguistics* (2nd ed.). (pp. 416–429). London: Routledge.

Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast. *Applied Corpus Linguistics*, *3*(3), 100066. doi:10.1016/j.acorp.2023.100066

Davies, M. (2008-) The corpus of contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/

Gilquin, G. (2016). Input-dependent L2 acquisition: Causative constructions in English as a foreign and second language. In De Knop, S., & Gilquin, G. (Eds.) *Applied Construction Grammar*. (pp. 115–148). Berlin: De Gruyter.

Gilquin, G. (2021). Using corpora to foster L2 construction learning: A data-driven learning experiment. *International Journal of Applied Linguistics*, *31*(2), 229–247. https://doi.org/10.1111/ijal.12317

Gilquin, G., & De Knop, S. (2016). Exploring L2 constructionist approach. In De Knop, S., & Gilquin, G. (Eds.). *Applied Construction Grammar*. (pp. 3–17). Berlin: De Gruyter.

Gilquin, G., & Granger, S. (2022). Using data-driven learning in language teaching. In O'Keeffe, A. & McCarthy, M. (Eds.) *The Routledge handbook of corpus linguistics* (2nd ed., pp. 430–442). London: Routledge.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at work. The nature of generalization in language*. Oxford: Oxford University Press.

Hamada, A., Iso, T., Kojima, M., Aizawa, K., Hoshino, Y., Sato, K., Sato, R., Chujo, J., & Yamauchi, Y. (2021). Development of a vocabulary size test for Japanese EFL learners using the new JACET list of 8,000 basic words. *JACET Journal*, *65*, 23–45.

Hilpert, M. (2019). *Construction grammar and its application to English* (2nd ed.). Edinburgh: Edinburgh University Press.

Hoffmann, T. (2022). *Construction grammar: The structure of English*. New York : Cambridge University Press.

Hoffmann, T., & Trousdale, G. (Eds.). (2013). *The Oxford handbook of construction grammar*. Oxford: Oxford University Press.

Iida, Y. (2021). Movie scripts for English learning/teaching material (2): On the way construction. *Annual report of the Faculty of Education, Gifu University: Humanities and social science*, *70*(1), 109–118.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *English Language Research Journal*, *4*, 1–16.

Leńko-Szymańska, A. & Boulton, A. (2015). *Multiple affordances of language corpora for data-driven learning. Amsterdam*: John Benjamins.

Luzondo Oyón, A. (2013). Revisiting Goldberg's semantic constraints on the 'way' construction. *Revista Española de Lingüística Aplicada*, *26*, 349–364.

Manabe, D. (2024). Testing the effectiveness of data-driven learning: Learning the way construction. *Human and Socio-Environment Studies*, *48*, 17–31.

Ministry of Education, Culture, Sports, Science and Technology. (2015). Kaku shiken dantai no dēta ni yoru CEFR to no taishōhyō [Comparison table with CEFR based on data from each testing organization]. https://www.mext.go.jp/b_menu/shingi/chousa/shotou/117/

shiryo/__icsFiles/afieldfile/2015/11/04/1363335_2.pdf

Mizumoto, A. (2023). Data-driven learning meets generative AI: Introducing the framework of metacognitive resource use. *Applied Corpus Linguistics*, *3*(3), 100074. https://doi.org/10.1016/j.acorp.2023.100074

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

O'Sullivan, Í. (2007). Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy. *ReCALL*, *19*(3), 269–286.

R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Sung, M., & Yang, H. (2016). Effects of construction-centered instruction on Korean students learning of English transitive resultative constructions. In De Knop, S., & Gilquin, G. (Eds.) *Applied Construction Grammar*. (pp. 89–113). Berlin: De Gruyter.

Takahashi, S., & Fujiwara, Y. (2016). Effects of inductive learning based on data-driven learning at elementary schools in Japan. *JES Journal*, *16*(1), 84–99.

Yoon, H., & Jo, J. W. (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.

（真部大輔　金沢大学大学院　Email: daisuke825@stu.kanazawa-u.ac.jp）

Appendix 1. The DDL material (the concordance)

Way 構文の説明と例文 1~20 を読み、ワークシート（別紙）を行ってください。

Way 構文

形：「主語＋動詞＋one’s way＋前置詞句/副詞」

意味：「動詞の表す行為をすることでできた経路を通り、前置詞句/副詞で示す方向へ主語が移動する（移動には困難や障害を伴うことが多い）。」

1	The officers **made their way through** the kitchen and into the lobby.
2	Unlucky for me, it seems that a mosquito **found its way into** my bedroom.
3	Perhaps he could still manage to **talk his way out of** this increasingly dangerous situation.
4	He rudely **elbowed his way through** the crowd toward her.
5	She is a frequent keynote speaker and radio show guest whose profound teachings have helped many **find their way through** the difficult times of life.
6	Jimmy will **find his way through** the dark forest.
7	The poem **found its way into** the pages of Punch magazine.
8	What is clear is that remarkably little of the agency's money **finds its way to** the people who need it.
9	He kept his hat on as he **made his way across** the living room and into the kitchen.
10	After eating breakfast, he **made his way through** the snow down the hill to where someone had a phone working and he called and had someone come plow us out.
11	Instead, she graduated, grabbed her tiny savings, and **made her way to** Nepal.
12	On Wednesday nights he drags the projector out of his office and sets it up in the art room and they watch Godzilla **making his way toward** Tokyo.
13	Instead, he decided to go to law school and **worked his way through** Yale Law School.
14	He **worked his way through** the crowd, toward the door.
15	We have to **work our way out of** this mess.
16	To **work your way into** a new community, where you're not very well known, you've got to be there at least 10 years and build all those relationships.
17	As she left him behind and **worked her way to** the opposite end of the crowd, she tried not to think about what the doctor had said.
18	The man left my side and, using his stick for aid, **pushed his way to** the front of the crowd.
19	He said he woke up smelling smoke and had to **fight his way out of** the burning building.
20	Several news crews **fight their way through** the crowd.

Appendix 2. The DDL material (the worksheet)

ワークシート

名前：____________________

例文（別紙）を見て、以下のタスクを行なってください。

1. 例文１と２を日本語に訳してください。

   例文１ ____

   例文２ ____

2. 例文３を読み、次の文を英語で言い換えてください。

   He **joked his way out** of difficult situations.

   ____

3. 例文４を英語で言い換えてください。

   例文４ ____

4. way 構文の形（品詞、動詞の種類など）について気づいたことを説明してください。

5. 例文を通して way 構文の意味についてどんなことがわかりましたか。動詞の意味に注目して way 構文の意味を説明してください。

6. way 構文について、他に気づいたことを記入してください。

Appendix 3. The translation tasks in the pre- and post-test (Due to space limitations, the answer sections were omitted)

事前テスト（the pre-test）

問. 以下の文を日本語訳してください。

(a) She made her way through the forest.

(b) They found their way to New York.

(c) My friend elbowed her way through the crowd.

(d) My sister talked her way out of the difficult situation.

(e) The student worked his way through high school.

事後テスト（the post-test）

問. 以下の文を日本語訳してください。

(a) He made his way through the crowd.

(b) She found her way to Tokyo.

(c) My brother elbowed his way through the crowd.

(d) I was able to talk my way out of the situation.

(e) My friend worked her way through university.

「論文」

エラータグ付き学習者コーパスを用いた英語ライティングの分析：中学・高等学校の学年進行に伴う特徴

阿部真理子

Abstract

This study investigates developmental patterns in the written English of Japanese EFL learners from the 7th to 12th grade by analyzing a corpus of approximately 30,000 words. The corpus was annotated with 37 types of error tags to examine error patterns and changes across different parts of speech and error type. The analysis revealed some findings. While overall error rates generally decreased with school year progression, there was a temporary increase in the ninth grade, third year of junior high school. This increase suggests that students at this stage attempt to use more complex linguistic features or newly learned grammatical structures. Regarding syntactic development, the analysis showed a clear development from simple to complex sentences, with senior high school students increasingly using more sophisticated structures involving conjunctions. These results provide insights into the characteristics of Japanese learners' linguistic development across different parts of speech and demonstrate qualitative changes in language use as students' progress through school years.

１．研究背景

学習者言語のエラー分析は，1970年代の第二言語習得研究において，中心的な研究手法として位置づけられていた。学習者の誤用を分析することで，言語習得のプロセスや学習者特有の困難点を明らかにすることができると想定されていた。Corder（1967）は学習者のエラーが単なる間違いではなく，言語発達のプロセスを示す重要な情報を提供していると主張した。当時はデータの収集や分析手法に限界があり，大規模かつ多角的な検証は困難であったが，1990

年代には情報処理技術の急速な進歩により，大量の電子化テキストを効率的に処理することが可能になった。これを背景に，学習者コーパス研究（Learner Corpus Research: LCR）という新たな研究分野が発展し始めた。LCR は，従来の研究では扱うことのできなかった規模の産出データ（学習者の書き言葉や話し言葉）を収集・電子化し，それらを計量的に分析することを可能にした。そして，このアプローチは，第二言語習得研究における仮説を実証的に検証するための新たな基盤となった。

国内における LCR には，日本語を母語とする英語学習者の書き言葉データを用いて形態素の習得順序研究を検証した Tono（2006）がある。研究の結果，日本語を母語とする英語学習者は冠詞の習得に困難を感じる一方, 所有格の “s” の習得は初期段階から比較的容易であることが明らかになった。Izumi and Isahara（2004）も，日本語を母語とする英語学習者の話し言葉データを用いて形態素の習得順序を検証し，Tono（2006）と同じ傾向が得られることを確認した。また形態素の習得順序には母語が影響を与える点や，特に冠詞の習得に困難が見られるという知見は，Murakami and Alexopoulou（2016）において示されている。

このような形態素の習得順序研究に加えて，LCR は学習者言語の特徴をより包括的に理解するために様々な言語項目に焦点を当てるようになった。コーパス言語学的な手法を用いて，学習者が言語発達の特定の段階において，どのような文法構造を使用しているのか, 過剰使用（overuse）や過少使用（underuse）を含めて調査された（e.g., Gilquin et al., 2008; Götz, 2015; Granger, 1998; Granger et al., 2013; Ishikawa, 2013）。また，学習者言語の発達段階を特定するための分析も進められ，異なる習熟レベルの学習者を区別するための言語項目が特定される研究が行われた。例えば，会話の流れを促進する表現（Hasselgren, 2001），語彙指標（Crossley et al., 2010a, 2010b, 2011），品詞連鎖（Tono, 2000），品詞（小林 , 2007b; Marsden & David, 2008），高頻度語（小林，2007a），現在時制の動詞，話者の評価や態度を表す副詞，機能語（小林，2010）などが挙げられる。

これらの研究は，従来の小規模な分析では捉えられなかった学習者言語の特徴を明らかにしてきた。コーパスデータを活用する主な利点は，言語処理技術に基づいた客観的な分析を行うことで，研究結果の信頼性が向上する点にある（MacWhinney, 2000）。一般公開されたデータを利用することにより，研究の再現性が確保され，研究者間でデータを共有できることも強みの一つである。さらにはコーパスデータを用いた量的研究と質的研究を組み合わせることで，よ

り一般化可能性の高い知見を得ることができる（e.g., Myles & Mitchell, 2004）。一方，様々な言語項目の分析に加えて，学習者言語に対する理解を深めるためには，学習者が産出する言語に含まれるエラーの特徴を詳細に分析する必要があると言える。

先行研究では，日本語を母語とする学習者の習熟度情報付き英語ライティングコーパスにエラー情報を付与し，学習者言語をエラーの観点から多角的に分析する研究が実施されている（Abe, 2007a, 2007b）。これにより，習熟度に応じて学習者言語のエラーパターンがどのように変化するのかが明らかになった。例えば，初級学習者は動詞関連のエラー（例：主語と動詞の一致）を，上級学習者は名詞関連のエラー（例：名詞の語彙選択）を多く産出する傾向があることが分かった。冠詞の正用率は，習熟度の発達とともに向上するものの他の言語項目と比較すると，日本語を母語とする英語学習者にとって，長期的に習得が困難な項目であることが示された。一方，動詞に関連する前置詞，主語と動詞の一致，アスペクト，名詞の屈折などのエラーは，学習が進むにつれて減少することが示された。

さらに阿部（2007c）は，日本語を母語とする英語学習者の英作文における誤用を，品詞，三つのエラータイプ（誤形成，脱落，冗長），習熟度の観点から分析した。第一に，誤形成エラーは正用率の高い品詞（例:代名詞，形容詞，副詞，名詞，動詞）と関連し，習熟度の向上とともに頻度が減少することが示された。第二に，脱落エラーは学習者が徐々に正しく使用できるようになる品詞（例：前置詞の補部や冠詞）と強い関連があり，変化のパターンは複雑であることが報告された。第三に，冗長エラーは主に接続詞と関連し，学習段階を通じて大きな変動は見られないが，習熟度が向上するにつれて徐々に減少する傾向が観察された。このようなカテゴリーごとの検証により，一部のエラーは習熟度の向上とともに減少する傾向があること，そして各エラータイプが習熟度において独特の変化パターンを示すことが明らかになった。特に，誤形成エラーの減少率，脱落エラーの複雑な変化パターン，および冗長エラーの緩やかな減少傾向は，学習者の言語発達段階を反映している。これらの特徴と習熟度との関係は，エラー分析が英語学習者の習熟度を客観的に区別し，発達段階を特定するための指標として機能することを示唆している。

近年では，学習者の言語使用におけるエラーを高度な統計処理技術を用いながら分析することで，どの言語項目がレベル判定に寄与するのかを明らかにする研究が進んでいる（石井・近藤，2015; Kobayashi, 2014; Kobayashi & Abe,

2016)。エラーの頻度と学習者の習熟度の情報を利用したこれらの先行研究では，冠詞や動詞の語彙選択，前置詞の語彙選択などがレベル評価に有効な項目として挙げられている。また Mizumoto and Watari（2023）は，Cambridge Learner Corpus First Certificate in English（CLC FCE）データセットを用いて，日本語を母語とする英語学習者に特有なエラーを分析し，限定詞，前置詞，名詞の三つに注目すべきであると指摘している。このような研究は，教育者や教材開発者，テスト開発者に貴重な情報を提供することができるだけではなく，自動採点システムの構築にあたり，より精度の高い評価を行うための基礎的な資料を提供することができる。

本研究では，日本語を母語とする中高生の英作文における品詞別・誤用タイプ別のエラー分析（阿部，2007c）を発展させ，学年進行に伴い英語ライティングがどのように変化するのか，そのプロセスを記述することを目的とする。阿部（2007c）ではエラー率の量的分析を中心としていたのに対し，本研究では文構造の複雑さや文法機能の変化に焦点を当て，質的側面から詳細に検証する。また，日本語使用の方略的な側面に注目し，これを単なるエラーではなく学習者言語の発達における指標の一つとして捉え直した。さらに，習得困難度を四つの枠組み（規則の内部構造の複雑さ，文法的機能の伝達の有無，母語との類似性，既習知識との関係）から解釈することで，中高生の英作文におけるエラーの特徴について，解釈を試みている。以上の背景を踏まえ，本研究では三つの観点から日本語を母語とする中高生の英語ライティングの実態解明を目指す。

1. 文構造の複雑さや文法機能の使用は，学年進行に伴いどのように変化するのか。
2. 日本語使用の方略的な側面はどのように変化するのか。
3. 日本語を母語とする英語学習者に特有の課題（冠詞，前置詞など）の発達過程には，どのような特徴があるのか。

2. エラータグ付き学習者コーパスの構築

2.1 分析データ

近年の AI 技術の飛躍的な進展により，学習者エラータグの自動付与が現実のものとなりつつある。この技術は，タグ付与に要する人的・時間的コストを

大幅に削減することが期待されている。例えば, Cambridge University Press から，英語学習者の誤用とその誤用を英語母語話者が訂正した情報が対になったデータが公開されており（Nicholls et al., 2024），学習者の原文と訂正文のデータから，誤用パターンを学習することが可能となっている。このような技術的進展を見据え，本研究では手動によって構築されたエラータグ付き学習者コーパスの分析を通じて，自動エラータグ付与の研究に資する知見の提供を目指す。

英語学習者のデータとして，Japanese EFL Learner（JEFLL）コーパスを利用した。JEFLL コーパスは，日本語を母語とする英語学習者の作文データを大規模に収集したものである（投野，2007）。本研究では，このコーパスから中高一貫校の学習者による英作文データを抽出し, エラータグを付与したサブ･コーパスを構築した。英作文のトピックは，JEFLL コーパスに収録されている 5 つのうち，「学園祭」に限定した。これはトピックによる言語使用の変動を最小限に抑えるためである。収集された英作文は，準備時間なし・参考資料なしという条件で 20 分間という制限時間内に書かれたものである。JEFLL コーパスの作文タスクでは，必要に応じて日本語の使用が認められており，日本語で書かれた箇所は <jp></jp> のタグで囲まれている。これにより，英語力の制約によって特定の表現が避けられるのを防ぎ，より自然な学習者言語データを収集することが可能となっている。サブ・コーパスの構築には，中学 1 年から高校 3 年までの各学年から約 5,000 語（誤差 5 語以内）を無作為に抽出し，合計 30,000 語規模のデータセットを作成した（表 1 参照）。

表 1. サブ・コーパスのサイズ

学年	人数	総語数	1 人あたりの平均語数	総語数標準偏差	総文数	1 文あたりの単語数	日本語の使用頻度
中学 1 年	104	4,994	48.02	13.52	722	6.92	782
中学 2 年	77	5,004	64.99	24.62	643	7.78	354
中学 3 年	87	5,000	57.47	22.07	623	8.03	364
高校 1 年	46	4,997	108.63	32.48	566	8.83	232
高校 2 年	53	5,000	94.34	60.80	541	9.24	244
高校 3 年	55	5,005	91.00	38.43	484	10.34	143
合計	422	30,000	71.09	40.19	3,579	8.38	2,119

2.2 エラータグ付与の方法

エラータグ付き学習者コーパスを構築するにあたり，本研究では汎用性を重視した。その汎用性を確保するため，National Institute of Information and Communications Technology Japanese Learner English（NICT JLE）コーパスで採用されているエラータグガイドライン（和泉 他，2004）を基盤とし，表 1 に示されたデータに対して表 2 のエラータグを付与した。ただし，中高生の英作文に特化したタグ体系とするため，以下の修正を加えた。第一に，前置詞の分類基準を変更した。NICT JLE コーパスでは「前置詞」（prp_lxc1）と「従属前置詞」（prp_lxc2）の区別があるが，本研究では「一般動詞以外の品詞と共起する前置詞」（prp_lxc1）と「一般動詞と共起する前置詞」（prp_lxc2）に分類した。第二に，スペリングエラーは頻度が高いため，その扱いを限定した。辞書に記載のない動詞の活用形（例：“make” の過去形として “maked” と記している場合）は，動詞の活用エラー（v_inf）として扱った（例：<v_inf crr=“made”>maked</v_inf>）。また，文頭の大文字と小文字の誤り，複合語のスペースの有無，本来 1 語または 2 語であるべき単語の誤り（例：basket ball）はエラーとして扱わなかった。第三に，日本語の使用に関連して，特殊なケースの処理基準を決めた。中学 1 年に多く見られた現象であるが，名詞を日本語で記述しながら英語の冠詞を付与するケースについては，分析の対象としていない。そして，誤って記された単語の品詞ではなく，正しいとされる品詞に基づいてエラータグを付与した。このタグ付与の方式は，NICT JLE のエラーガイドラインに準拠したものであり，Ellis（2005）においても言及されている方式である。例えば “afraided” という誤用は，正用が “afraid” であることから形容詞のエラー（aj_lxc）としてタグ付けを行った（例：<aj_lxc crr=“afraid”> afraided</aj_lxc>）。

次に信頼性の高いタグ付与を実現するために必要である明確な基準と手順を策定した。タグ付与基準として, 各エラータイプの判定基準を具体的に決定し, 複数のエラーが重複する場合の優先ルールを設けた。タグ付与の信頼性を確保するため, 三つの原則を設定した。第一の原則として，文法項目に焦点を当て，スタイルやディスコースレベルのエラーは対象外とした。第二の原則では，最小限の修正で文法的に正しい形に修正可能な候補を挙げることとした。第三の原則として，明確な誤りのみを対象とし，判断が分かれるケースは除外する方針とした。タグ付与の手順としては，中学校・高等学校で教育経験のある教員 1 名が，四つのステップを踏みながら実施した。第一段階として，500 語のサ

表 2. エラータグ一覧表（和泉 他，2004 を改変）

品詞	エラーの種類	タグ	エラーの例
名詞 [n]	1. 活用	[n_inf]	*childerens / *housewifes / *peoples
	2. 単複	[n_num]	many *kind of
	3. 格	[n_cs]	my *friend house
	4. 可算/不可算	[n_cnt]	*a music / *musics
	5. 語彙選択	[n_lxc]	a *type (a typewriter) / *catch ball
動詞 [v]	1. 活用	[v_inf]	*sleeped
	2. 主語動詞の人称・数の不一致	[v_agr]	And many of them *was crying.
	3. 形の選択	[v_fml]	I’m already *expect to the next festival so much.
	4. 時制	[v_tns]	I *eat breakfast this morning.
	5. 相	[v_asp]	The people *weren’t knowing the reality.
	6. 態	[v_vo]	We *were played very well at the festival.
	7. 定形・不定形の選択	[v_fin]	*Prease* *visiting us school.
	8. 否定形	[v_ng]	*not well
	9. 疑問形	[v_qst]	think about *how should we use the room
	10. 補部	[v_cmp]	The advertisement makes people *to understand how dangerous drugs can be.
	11. 語彙選択	[v_lxc]	She *is black and short hair.
形容詞 [aj]	1. 活用	[aj_inf]	*more tall
	2. 原形・比較級・最上級の用法	[aj_us]	Jane is taller than Mary, but Mary is the *best basket ball player.
	3. 形容詞の数	[aj_num]	She worked hard to help the *poors.
	4. (修飾語として) 数量を表す形容詞	[aj_qnt]	There was *few traffic on the road.
	5. 補部	[aj_cmp]	It was kind of you *helping him.
	6. 語彙選択	[aj_lxc]	It is a *genius diamond.
副詞 [av]	1. 活用	[av_inf]	*more far
	2. 原形・比較級・最上級の用法	[av_us]	She came back *most quickly than me.
	3. 副詞の位置	[av_pst]	I have difficulty *often in understanding her.
	4. 語彙選択	[av_lxc]	He worked *hardly today.
前置詞 [prp]	1. 補部	[prp_cmp]	I look forward *to see you again.
	2. 語彙選択誤り 1	[prp_lxc1]	It was held *on June.
	3. 語彙選択誤り 2	[prp_lxc2]	Tom’s teacher accused him *about cheating.

表 2. つづき

冠詞 [at]	1. 冠詞	[at]	It was * wonderful day.
代名詞 [pn]	1. 活用	[pn_inf]	*themselfes
	2. 数・性別の一致	[pn_agr]	It is a good book. I like *them.
	3. 格	[pn_cs]	*We school festival is very good.
	4. 語彙選択	[pn_lxc]	I was in Brass Band, *it played music in big hole.
接続詞 [con]	1. 語彙選択	[con_lxc]	Some of <jp>お客さん</jp> moved and bagan to cry, * they said "That was so nice!!"
関係詞 [rel]	1. 格	[rel_cs]	She is the girl *who smartphone was stolen.
	2. 語彙選択	[rel_lxc]	The cake * is on the table looks delicious.

ンプルデータを用いた試行を行い，タグ付与基準の確認と調整を行った。第二段階では，全データに対して1回目のエラータグ付与を実施した。第三段階では，1週間の時間的な間隔を置いた後，2回目のタグ付与を行った。最終段階では, 1回目と2回目で不一致が見られた箇所を中心に確認作業を行い, エラータグを確定した。この一連の作業過程においては，タグ付与中に発見された基準の改善点を記録し，作業に反映した。なお，本研究で使用したデータは，前節で述べたようにJEFLLコーパスから抽出されたもので，各学年の異なる学習者から収集された横断的（cross-sectional）なものであり，同一の学習者を追跡した縦断的（longitudinal）なデータではない。よって，本研究の結果は，集団としての傾向を示すものであり，個人内の発達過程を示すものではない。

これらの手順と確認プロセスにより，エラータグ付与の一貫性と再現性を向上させることを目指した。図1は，本研究におけるエラータグ付与の具体例を示している。次に，以下の例文に含まれる時制のエラー（Our class <v_tns crr="made">make</v_tns> a piramid's inside.）を用いて，エラーの構造と意味を説明する。このエラーには，<v_tns></v_tns>（動詞の時制エラー）のタグを付与し，修正候補として crr="made" が記されている。これは，過去に行われた学園祭について述べられているため，動詞の過去形 "made" を使用すべきところに現在形 "make" が誤用されていることを示す。なお, この例文中の "piramid" は学習者によるスペリングの誤りであるが，スペリングエラーは分析対象外としているため，エラータグは付与していない。

<s>We held a school festival <prp_lxc1 crr="on"></prp_lxc1> September 14 and 15.</s> <s>Our class <v_tns crr="made">make</v_tns> a piramid's inside.</s> <s>It <v_tns crr="was">is</v_tns> very dark, and almost a ghost house.

注．<s>は文章の始まりを，</s>は文章の終わりを示す。

図 1. エラータグが付与されたデータの例

2.3 分析

エラー分析に加え，正用法の比率を算出するために Constituent Likelihood Automatic Word-tagging System（CLAWS）7 の品詞タグを活用した。ただし，学習者の誤用により品詞タグが正しく付与されない場合や，国内で指導されている学校文法との違いが見られる場合には，手作業で修正を行った。分析では，まず品詞別のエラー頻度を確認した。具体的には，名詞，動詞，形容詞，副詞，前置詞，冠詞，代名詞，接続詞について，エラーの総数を集計した。その上で，各品詞における使用総数（正用と誤用の合計）に対するエラー率を算出し，学年進行に伴うエラー率の変化を確認した。また，文の構造面では，単文から重文と複文への発達過程を確認するために，特に接続詞の使用パターンと文の結合方法を分析した。これらを通じて，品詞，文法といった多様な言語項目に注目し，学年ごとの学習者データを比較した。これにより，各学年に特徴的なエラーや，すべての学年に共通するエラーのパターンを確認することができた。また，各学年における日本語タグ（<jp>）の出現頻度を集計するとともに，日本語使用のパターンを分類した。

3. 結果

3.1 学年進行によるエラー率の変化

全体的なエラー率は，中学 1 年で 9.2% と最も高かった。中学 2 年では 6.9% まで減少したが，中学 3 年では 8.1% と再び上昇した。その後，高校 1 年で 7.7%，高校 2 年で 5.3% と減少傾向を示し，高校 3 年では多少上昇して 5.7% となった。この結果から，全体的な傾向としては学年が進行するにつれ，エラー率が低下するが，中学 3 年での一時的な上昇が特徴的な変化として観察された。図 2 に示されるように，この上昇はこの時期に最も高いピークを示す冠詞のエラー率（42.1%）が影響していると考えられる。中学3年では，前置詞のエラー率（21.3%）と接続詞のエラー率（13.4%）も，それぞれ上昇している（図 2 参照）。この

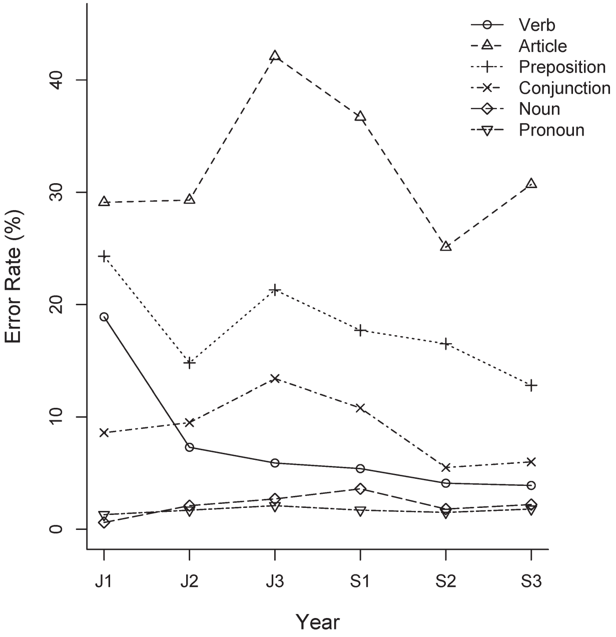


図 2. 学年進行による品詞別エラー率の変化

ような結果から，特に中学 2 年から中学 3 年への移行期において，品詞の特性に応じた指導上の配慮が必要であることが示唆される。

3.2 品詞別エラー率の特徴

品詞別のエラー率の高さによって，品詞を 3 つのグループに分類できる。一つ目は，高エラー率のグループ（冠詞 25.1〜42.1%，前置詞 12.8〜24.3%）である。二つ目は，中エラー率のグループ（接続詞 5.5〜13.4%，動詞 3.9〜18.9%）である。三つ目は，低エラー率のグループ（名詞 0.6〜3.6%, 代名詞 1.3〜2.1%）である。品詞別の特徴としては，冠詞のエラー率は他の品詞と比較した場合，全学年を通じて最も高い。前置詞は，特に中学 1 年と中学 3 年において高い。接続詞は，中学 3 年でピークとなるが，それ以降は下降する傾向にある。一方，動詞，名詞，代名詞は低いエラー率を示しており，学年進行による大きな変化は見られない。この品詞別のエラー率の特徴は，CLC FCE データセットを利用した最近の学習者コーパス研究の結果（Mizumoto & Watari,

2023）とも一致する。

3.3 日本語使用の特徴と変化

本研究では，JEFLL コーパスにおける日本語の使用をエラーではなく，学習者言語の発達における方略的な言語使用として捉えた。母語の使用は学習者の言語使用を引き出し，より豊かな表現を可能にすると考えられる。サブ・コーパス全体の総文数 3,579 文のうち，2,119 件の日本語の使用が確認された。ただし，一文中に複数回，日本語が使用されているケースもある（例：<jp> 当日は </jp> many <jp> 人がきてました </jp>）。中学 1 年における日本語の使用は 782 件であったが，学年が上がるにつれて減少し，高校 3 年では 143 件となり，段階的な変化が確認された。分析の結果から，日本語の使用にも三つのパターンが特定された。第一に，文全体の日本語使用である（例：「おかしについての展示をしました」）。第二に，学園祭に特有の具体的な物の名称を示す日本語（例：「紙芝居」，「ダンボール」，「木材」，「忍者屋敷」）が，英語における語彙知識の不足を補う目的で使用されている。第三に，助詞や文末表現の日本語使用であった（例：「は」，「が」，「でした」）。そして，これらの日本語使用パターンは，学年に応じて異なる特徴を示した。例えば，中学 1 年では文全体の日本語使用が特徴的であった。中学生の時期は，基本的な名詞や動詞の日本語使用が中心であり，英文中に部分的に組み込まれていた。しかしながら特に高校 2 年では，他の学年と比べて特徴的な日本語使用のパターンが見られた。それは微妙な感情のニュアンスを示す日本語（例：「ぶっちゃけて」，「しらけてる」）の使用，および複雑な状況や動作を説明する日本語（例：「ほきょうしている」，「ビンタした」）の使用であった。高校 2 年における日本語の使用においては，感情表現が日本語使用全体の 10.7% を占め，複雑な状況説明に関する表現は 9.9% であった。さらには，高校 2 年と高校 3 年の間において質的な変化が見られた。高校 2 年では感情表現として基本的な語彙（happy, glad：11 件）が使用されているのに対し，高校 3 年になるとより洗練された語彙（satisfied, delighted, delightful, pleased：9 件）が使用されている。これは感情を英語で表現しようとする意欲の表れと解釈でき，それに伴い日本語使用が減少すると考えられる。また，高校 2 年と高校 3 年の間には，日本語使用の質的な違いだけでなく，量的な違いもあった。1 人あたりの日本語の使用回数は高校 2 年では 4.6 回であったが，高校 3 年では 2.6 回に減少している。高校 3 年では，日常的な感情は英語で表現し，日本語は文化祭に特有の名詞（例：“Anmitsu,” “Yakisoba,” “Kanten”）に限定される傾向があった。このような日本語使用の変化は，学習者言語の発達を反映していると考えられる。初期段階では，学園祭の内容を伝達するための手段として日本語に依存する傾向があったが，学年の進行に伴い日本語の使用が減少し，より戦略的に日本語を使用する段階へと移行していたと言える。

4. 考察

4.1 高エラー率の品詞（冠詞，前置詞）

3.2 節で示された品詞間でのエラー率の違いは，各品詞の特性によって説明できる。特に，日本語に対応する概念がない品詞（冠詞，前置詞）は，エラー率が高い傾向があった。これらの特徴をより詳細に分析するため，この節では文法項目の習得困難度を説明する枠組みを使って解釈を試みる。まずは，3.1 節で示した中学 3 年での冠詞エラー率の上昇（42.1%）について触れる。この急激な上昇は，U 字型発達曲線（e.g., Kellerman, 1985）の特徴と一致している。U 字型発達曲線とは，学習者が新しい言語規則を取り入れる過程で一時的にエラー率が上昇し，その後熟達とともに再び減少するというパターンを指す。この現象は単に新しい文法項目への試行による結果とは解釈できない。むしろ，この時期に学習者が文章の複雑化を試みる中で冠詞の使用機会が増加し，その結果としてエラーも増加したと考えられる。

冠詞と前置詞の高いエラー率に関しては，白畑（2016）に示されている四つの観点（規則の内部構造の複雑さ，文法的機能の伝達の有無，母語との類似性，既習知識との関係）から説明できる。第一に，規則の内部構造の複雑さの観点からすると，冠詞は定冠詞・不定冠詞の選択，可算・不可算の区別など，複数の判断基準を同時に考慮する必要がある。前置詞に関しては，1 つの前置詞が文脈に応じて多様な意味を持つために学習の混乱を招きやすい（例：空間 in the box，時間 in winter，抽象的状態 in anger，慣用的表現 in fact）。また，語連鎖の構成語として多義性を帯びる場合（例：bring on, hold on）があるだけではなく，形容詞との組み合わせ（例：good at, interested in）も学習しなければならない。

冠詞と前置詞のエラー率の変動には，学習者の段階的な発達が影響しているとも考えられる。まず冠詞に関しては，中学 1 年と中学 2 年では基本的な使用（例：a book）に留まっている。しかし中学 3 年になると，既知と未知の概念（例：

Our class did a play whose name is <jp> 漂流教室 </jp>. The play is parody of TV.）や総称的なものを示す用法（例：Making <jp> 看板 </jp> is most difficult!）など，より複雑な判断を要する用法が使用されるため，エラー率の上昇が引き起こされると考えられる。前置詞に関しては，中学 1 年から中学 2 年にかけては，基本的な空間を示すための表現（例：in the classroom）が中心であるが，それ以降はより複雑な前置詞の用法も使用されるようになる。高校 1 年以降においてはこれらの用法に関する理解が徐々に定着するため，エラー率が低下すると考えられる。

第二に，文法的機能の伝達の観点からすると，両品詞とも具体的な意味内容よりも文法的機能としての役割を果たしている。冠詞の場合，新情報と既知情報の区別を示したり, 普遍的な対象（例：the earth）であることを示したりする。前置詞に関しても，名詞句内の要素の関係性（例：the book on the desk）を示したり，動作の方向性や手段（例：I go to school by bus.）を示したりする。これらの役割は，具体的な物を表す名詞や動作を表す動詞とは異なり，文の要素の関係性を示す。そのため，個々の意味を覚えるだけではなく，文脈を通して他の語句とどのように関係づけられるかを理解する必要があり，結果的に冠詞と前置詞のエラー率が高くなると考えられる。

第三に，母語との類似性の観点からすると，冠詞は日本語に対応する文法上のカテゴリーが存在しない。前置詞も日本語の助詞と一対一の対応をしていない（例：学校に行く go to school，机の上に置く put on the desk，3 時に at 3）。第四に，既習知識との関係からすると，冠詞は規則を学習した後も文脈による判断が必要になるため，既習知識を単純に利用できない。例えば “I saw a dog yesterday. The dog was chasing a cat.” では新情報として不定冠詞を使用し，二文目では既に言及されている犬に対して定冠詞を使用する必要がある。また, “She goes to school every day.” という行為を表す場合は無冠詞でよいが，“She goes to the school to meet her teacher.” という例のように特定の学校を指す場合は定冠詞が必要である。さらには，“Water is essential for our life.” というように一般的な水を指す場合は無冠詞だが，“The water in this bottle is clean.” という例のように特定の水を指す場合は定冠詞を使う。このように文脈によって使うべき冠詞が変わるため，学習者は単純に規則を適用するだけでは不十分であり，文脈に応じた判断が求められる。また前置詞に関しては，同じ前置詞が異なる意味を持つことも多い。例えば，前置詞 “on” に関しては多くの学習者にとって接触の意味を示す用法（例：put on）は馴染みが深いが，継続の意味を示す用法（例：

move on）は新規性が高いと考えられる。このように冠詞と前置詞は全ての観点において，習得の難しさを示している。

4.2 中エラー率の品詞（接続詞）

接続詞の特徴について，文構造が複雑化する学年の進行段階に沿って説明する。第一段階（中学 1 年）は，単文に依存する時期である。この段階では，一文一文が独立しており，文と文のつながりが意識されていない。第二段階（中学 2 年）は並列構造が出現する時期である。名詞句の並列（例：We enjoyed games, food, and music.）が特徴的であり，等位接続詞を用いた単文の接続が試みられるようになる。第三段階（中学 3 年）はエラー率が最も高くなり，文の接続を試行錯誤する時期であると考えられる。接続詞を使用せずコンマで文をつなげる傾向や，文の構造上適切ではない接続の方法（例：We make smile and my <jp>“Irassyaimase”</jp> or <jp>“Arigatougozaimashita”</jp>.）が観察された。このような例は，接続詞の使用と文構造の理解が深まっていないことを示している。

第四段階（高校 1 年）では依然としてエラー率が高いものの，接続詞を用いた重文と複文が使用されるようになっている。1 文あたりの接続詞の使用件数は, 中学 3 年から高校 1 年にかけて多少の変化が確認された（中学 3 年：0.4 件，高校 1 年：0.6 件）。また “But,” “Because,” “So” などの接続詞を文頭に置く用法（例：Our class had a casino. But I couldn’t join the casino. Because I had a tennis match that day.）が多用されており，文頭に接続詞を置く文の頻度は中学 3 年の 19.5% から高校 1 年の 25.1% へと増加している。接続詞を種類別に分析すると，等位接続詞の使用頻度は，中学 3 年の 217 件から高校 1 年の 282 件に増加している。特に “and” の使用は 94 件から 143 件へと増加し，複数の節を連結しようとする傾向が強まっていると考えられる。同様に，従属接続詞の使用も 48 件から 63 件へと増加している。さらに接続詞の使用が量的に増えるだけではなく，質的な変化が生じている。特に “and” を用いた単純な連結が増加する一方，“because” や “if” などの従属接続詞の使用が見られるようになり，単に文を長くするだけでなく，因果関係や条件関係といったより複雑な論理関係を表現しようとする傾向がある。

第五段階（高校 2 年）になると，エラー率が低下すると同時に，複文構造の発達が見られる。単文から接続詞や関係詞を用いた複文や重文へと発展する傾向がある（例：There are many games using our muscle and 競う who is the best.）。

より複雑な従属節の使用が増加し，統語的複雑さが発達しており，時間や原因を表す副詞節の使用も確認できる（例：I don't know what wasn't good, so I'm 悔しい）。また，文と文のつながりに関する意識が高まり，“However,” “But,” “So,” “Because” などの接続詞を用いて理由を説明し，論理的な文の展開を試みる例が見られる（例：We couldn't often? the *survece* to all of them. Because of the time limit.）。

第六段階（高校 3 年）は文構造が成熟する時期であると言える。複数の節を組み合わせた表現（例：On the first day, when only the students enjoy other classes' exhibition, I had a lot of fun with the matches.）や，名詞句を主語とする用法（例：What we had done was a food shop at one of the festival in this school.），過去分詞による後置修飾を含む複雑な名詞句（例：The most exciting game was the one played by a girl named Momoko and a boy.）など，より高度な構文が使用されている。そして，高度な接続表現（例：as a result）の使用が可能となり，より論理的な文章構造が確立されている。このような発達は，文の長さや複雑さの向上だけではなく，洗練された思考を表現できるプロセスも示している。なお，関係詞節の使用においては，関係代名詞の完全な脱落（例：Some of my classmates sang pop songs and danced on the stage we made in the dark class room.）があったり，複雑な構文の使用に伴い，基本的なエラーが再出現したりしている。また，関係代名詞と代名詞の重複使用（例：My class did a show that we sang and danced in it.），誤った非制限用法の関係詞節（例：Some boys including student in a lower year took part in the game, which intensity surprised me.）といったエラーが観察された。

4.3 中エラー率の品詞（動詞）

動詞と時制の使用に関しても，学年進行に合わせて発達段階を説明することができる。第一段階（中学 1 年）は be 動詞に依存する時期であり，二つのエラーパターンが観察された。一つは「be 動詞 + 動詞の原形」（例：Festival is go away.）であり，もう一つは一般動詞の代わりに be 動詞を使用するパターン（例：I am a 受付）であった。また，“It's” を過去時制の文脈でも使用するなど，時制の使い分けへの意識が低い。第二段階（中学 2 年）は，一般動詞の使用が拡大する時期であると同時に，動詞に続く前置詞の脱落が観察された。第三段階（中学 3 年）は動詞の構造が複雑化する時期であり，「一般動詞＋一般動詞」という誤った連鎖が観察された。また自動詞・他動詞の区別が曖昧であるため，前

置詞の不必要な挿入（例：discuss about the festival）も観察された。

第四段階（高校 1 年）は，受動態の使用を試行錯誤する時期であり，“were enjoy” から “were enjoyed” への変化が見られたが，能動態を使用すべき文脈での不適切な受動態の使用（例：That was a good festival because guests are enjoyed.）も観察された。受動態の形式的な理解は進んでいるが，適切な文脈での使用は発達途上にあることが示唆される。また，時制の使用が複雑化する様子が観察された。時制の操作に関しては発達途上にあり，完了形や過去進行形などの使用が十分に定着していない（例：I hadn't played when we practiced, but I could play on the stage.）。

第五段階（高校 2 年）は時制と相の使用が拡張される時期であり，完了形や仮定法（例：If I had come to this festival, I wouldn't have come a school.）など，より複雑な動詞形式の使用が試みられている。ただし，不必要な過去完了形を使用する例も観察されている（例：I bought about ten books and I had read it all day.）。また，過去の出来事を描写する際に，時制を一貫させることに苦心していることが示唆された（例：We practiced really hard for the school festival to play well. Many people came to our concert. I am so tired after it...）。特に客観的な事実と主観的な感想を同時に述べる場合に混乱しているようでもある（例：Our school festival is held on September 14th and 15th. The first day, only students can enjoy the festival. I think it was not necessary.）。これらの例は，時制と相の習得がまだ発達段階にあることを示す特徴と言える。

第六段階（高校 3 年）は，動詞が多様化する一方，受動態の使用が依然として課題となっており，基本的なエラーが再出現する時期であると言える。例えば，“A number of guests bought the sweets, our shop was flourish.” のような文では，“flourish” という洗練された動詞を使用しようとする過程で，本来は能動態で使うべき自動詞を誤って受動態にしたため，文法的に不適切な表現となっている。これは，より洗練された動詞を用いる際には，適切な態を選択することが難しいことを示唆している。高校 3 年のデータにおいて，態に関するエラーは 4 件検出されている。具体的には，“Our school festival hold on 14th, 15th September.” のように be 動詞が欠落するケースや , “Our festival's symbol is changed every year.” のように能動態で表現すべきときに受動態を用いるケースなどが観察された。受動態に関するエラーは学年進行に伴い，件数が減少しているものの（高校 1 年：12 件 , 高校 2 年：5 件，高校 3 年：4 件），態の選択に関する課題が残っていることを示している。これらの現象は，より洗練された語彙を使

用する際に，文法的な正確性よりも意味の伝達を優先した結果，基本的な文法規則への注意力が低下したと考えられる。また語彙や表現の複雑化は，記述の内容にも影響を与えている。中学1年から2年では単純な事実描写が中心であり，時系列に沿った記述が特徴的であったが，中学3年から高校1年になるとより詳細な状況説明や因果関係の記述を試みることが多くなっている。

5. 結論

本研究は，日本語を母語とする中高生の英語ライティングにおけるエラー分析を行うことによって各学年における言語使用および発達上の特徴を明らかにした。また，各学年に特有のエラーパターンを報告した。これらの発見は，英語指導において発達段階に応じた指導方法の重要性を示唆するものでもある。結果として，具体的な意味内容を持つ言語項目（例：名詞，代名詞）は，中高6年間を通してエラー率が低かった一方，文法的な関係性を示す機能を持つ言語項目（例：冠詞，前置詞）は，規則の内部構造の複雑さ，文法的機能の伝達の有無，母語との類似性の低さ，既習知識との関係など，複数の要因に影響されることが示唆された。また学年進行に伴い，文構造の複雑化とともに使用できる表現の幅が広がっていくことも確認された。今後の課題として，作文のトピックやタスクの種類により分析結果に違いが出るのか，また異なる学校で得られたデータを用いた場合も同様の結果が得られるか，さらにどのような要因（例：母語，教科書，学習時間や方法）が学習者の英作文の発達過程に影響しているのかなどについても研究が必要となるだろう。

謝辞

本稿の執筆にあたり，貴重なご意見とご助言をいただいたすべての先生方に深く感謝申し上げます。また本研究に対し，多大なる支援をいただいたすべての方に感謝の意を表します。

参考文献

Abe, M. (2007a). A corpus-based investigation of errors across proficiency levels in L2 spoken production. *JACET Journal*, *44*, 1–14.

Abe, M. (2007b). Grammatical errors across proficiency levels in L2 spoken and written English. *The Economic Journal of Takasaki City University of Economics*, *49*, 117–129.

阿部真理子（2007c）．「JEFLL コーパスに見る品詞別エラーの全体像」投野由紀夫（編）『日本人中高生一万人の英語コーパス “JEFLL Corpus”—中高生が書く英文の実態とその分析—』（pp. 146–158）．小学館.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press. https://doi.org/10.1017/CBO9780511804489

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics*, *5*, 161–169.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010a). The development of polysemy and frequency use in English second language speakers. *Language Learning*, *60*(3), 573–605. https://doi.org/10.1111/j.1467-9922.2010.00568.x

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010b). The development of semantic relations in second language speakers: A case for Latent Semantic Analysis. *Vigo International Journal of Applied Linguistics*, *7*, 55–74.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*(2), 243–263. https://doi.org/10.1177/0265532211419331

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford University Press.

Gilquin, G., Papp, S., & Díez-Bedmar, M. (2008). Introduction. In G. Gilquin, S. Papp & M. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. vii–xi). Rodopi.

Götz, S. (2015). Tense and aspect errors in spoken learner English: Implications for language testing and assessment. In M. Callies & S. Götz. (Eds.), *Learner corpora in language testing and assessment* (pp. 191–216). Benjamins.

Granger, S. (1998). The computerized learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3–18). Addison Wesley Longman.

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2013). Twenty years of learner corpus research: Looking back, moving ahead. *Proceedings of the first learner corpus research conference (LCR 2011)*.

Hasselgren, A. (2002). Learner corpora and language testing: Small words as markers of learner fluency. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 143–173). Benjamins.

Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge University Press.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world* (Vol. 1, pp. 91–118). School of Language and Communication, Kobe University.

石井雄隆・近藤悠介（2015）．「文書分類の手法と一般化線形モデルを用いた英語ライティングにおける文法的誤りの影響」『人文科学とコンピュータシンポジウム』，71–76.

Izumi, E., & Isahara, H. (2004). Investigation into language learners' acquisition order based on an error analysis of a learner corpus. *Proceedings of the Interactive Workshop on Language e-Learning* (IWLeL 2004) (pp. 63–71).

和泉絵美･内元清貴･井佐原均（編著）（2004）．『日本人1200人の英語スピーキングコーパス』アルク．

Kellerman, E. (1985). If at first you do succeed. In S. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 345–353). Newbury House.

小林雄一郎（2007a）．「The NICT JLE corpus と語彙研究」『英文学誌』，*49*，17–29.

小林雄一郎（2007b）．「The NICT JLE Corpus における発達指標の研究―コレスポンデンス分析によるタグ頻度解析」『言語処理学会第13回年次大会発表論文集』，486–489.

小林雄一郎（2010）．「コレスポンデンス分析：データ間の構造を整理する」石川慎一郎・前田忠彦・山崎誠（編）『言語研究のための統計入門』（pp. 245–264）．くろしお出版.

Kobayashi, Y. (2014). Computer-aided error analysis of L2 spoken English: A data mining approach. *Proceedings of the Conference on Language and Technology 2014*, 127–134.

Kobayashi, Y., & Abe, M. (2016). Automated scoring of L2 spoken English with random forests. *Journal of Pan-Pacific Association of Applied Linguistics*, *20*(1), 55–73.

MacWhinney, B. (2000). *The CHILDS project: Tools for analyzing talk*. Erlbaum.

Marsden, E., & David, A. (2008). Vocabulary use during conversation: A cross-sectional study of development from year 9 to year 13 among learners of Spanish and French. *Language Learning Journal*, *3*(2), 181–198. https://doi.org/10.1080/09571730802390031

Mizumoto, A., & Watari, Y. (2023). Identifying key grammatical errors of Japanese English as a foreign language learners in a learner corpus: Toward focused grammar instruction with data-driven learning. *Asia Pacific Journal of Corpus Research*, *4*(1), 25–42. https://doi.org/10.22925/apjcr.2023.4.1.25

Murakami, A., & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*, *38*(3), 365–401. https://doi.org/10.1017/S0272263115000352

Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research*, *21*(4), 373–391. https://doi.org/10.1191/0267658305sr252oa

Myles, F., & Mitchell, R. (2004). Using information technology to support empirical SLA research. *Journal of Applied Linguistics*, *1*(2), 169–196.

Nicholls, D., Caines, A., & Buttery, P. (2024). *The write & improve corpus 2024: Error-annotated and CEFR-labelled essays by learners of English.* Apollo - University of Cambridge Repository. https://doi.org/10.17863/CAM.112997

Pendar, N., & Chapelle, C. (2008). Investigating the promise of learner corpora: Methodological issues. *CALICO Journal*, *25*(2), 189–206.

白畑知彦（2016）.「明示的文法指導，明示的フィードバックが効果的な文法項目とそうでない文法項目―項目別に教え方を変えてみよう―」外国語教育メディア学会関西支部 2016 年度春季研究大会.

Tono, Y. (2000). A corpus-based analysis of interlanguage development: Analyzing part-of-speech tag sequences of EFL learner corpora. In B. Lewandowska-Tomaszczyk & P. J. Melia (Eds.), *PALC'99: Practical applications in language corpora* (pp. 323–340). Peter Lang.

Tono, Y. (2006). L2 acquisition of grammatical morphemes. In T. McEnery, R. Xiao & Y. Tono, *Corpus-based language studies: An advanced resource book* (pp. 247–263). Routledge.

投野由紀夫（編著）（2007）.『日本人中高生１万人の英語コーパス―中高生が書く英文の実態とその分析』小学館.

（阿部真理子　岡山大学　Email: abema@okayama-u.ac.jp）

「論文」

HAVE the advantage が従える同格節 that —that 節中の助動詞 can は必須要素か—*

土屋　知洋

Abstract

The main aim of this paper is to investigate whether the abstract noun *advantage* can be followed by an appositive *that*-clause and whether the auxiliary verb *can* is essential within the clause to explain *advantage*. A descriptive study using both quantitative and qualitative approaches, incorporating corpus analysis and informant surveys, reveals two key findings. First, constructions in which *advantage* is followed by an appositive *that*-clause are observed in the spoken British English as well as in written forms of both British and American English. Second, the auxiliary verb *can* in such clauses is not a mandatory component for the noun *advantage* but is context-dependent. Its occurrence is strongly influenced by the meanings of “ability” or “circumstantial ability” conveyed in the appositive clause. These findings offer valuable insights for revising the treatment of *advantage* in reference books and dictionaries.

1. はじめに

（1）は, 名詞 advantage の成句として HAVE the advantage of（～という強み［長所］を持つ）をあげる海外と日本の学習辞典にみられる用例である（以下，イタリック，下線，そして波線は全て著者による）。[1]

（1）a. She *had the advantage of* a good education.　　（OALD10, 2020, “advantage”）
　　b. Ken *has the advantage of* speaking Italian.　　（W4, 2018, “advantage”）
　　c. The equipment *has the* additional *advantage of* being easy to carry.
　　　　（MED2, 2007, “advantage）

一般的に，上記（1）の用例が示すように（1c）の形容詞で修飾された若干のバリエーションは見られるものの，HAVE the advantage of の後ろには名詞句や動名詞を従える点で共通している。しかしながら，イギリス人ネイティブスピーカーは（3）の理由で，（1b）（1c）と同形式の動名詞を従える（2a）の英文は正しくなく，（2b）が正しいという。概略，形式的な文において，長所や強みを表すのであれば“能力”を表す can を含めたほうがより良いとして，(2b) の形式が適切だという主張である。

（2）a. ? ... online classes *have the advantage of* being taken at any time.
　　b. ... online classes *have the advantage that* they can be taken at any time.
（3）I think the first version might be accepted in a casual email, but for exam purposes it is not correct. Because an advantage is being described, it is better to include “can”.（You can do this if you want.）Therefore I think the second version is correct.

ここで問題となるのが，本稿で焦点を当てる成句 HAVE the advantage of において，各種辞典や参考文献に名詞 advantage が前置詞 of の代わりに同格を表す that 節を従えることに触れるものが皆無に等しく，そもそもこの用法が可能なのか，更にイギリス人ネイティブスピーカーの指摘する同格節内に“能力”を表す助動詞 CAN が必要なのか，という 2 点である。本稿では，この 2 点の語法的問題について，成句として確立している HAVE the advantage of が従える語彙要素と比較しながら，その生起条件についても論じる。まず，第 2 節で数少ない先行研究を概観し，第 3 節と第 4 節でコーパスのデータに基づき量的・質的観点から実証的に分析する。そして，最終の第 5 節で分析結果の妥当性を確認すべくインフォーマント調査を行い，（2b）の用法の可能性を示すと共に助動詞 CAN の生起が同格 that 節内の命題の表す意味と密接に関係していることを実証する。

2. 先行研究

この第 2 節では，まず名詞 advantage が同格節を導くことが可能なのか，成句表現の of の正体を探りながら確認する。また，仮に同格の that 節を導くことが可能であれば，その同格節内に助動詞 CAN が必要なのかについても先行

研究を概観する。

2.1 advantage が従える同格 of

同格を表す後置修飾には，以下（4）の 3 つのパタンが存在する。実際，Biber et al.（1999，p. 654）では，用例は示されていないが advantage を同格節の説明の一部で「抽象名詞 + of + ing」のリストに含めている。また，（5）の辞典では《◆ of は同格の of》と注付きで用例を上げている。

（4）［1］to-infinitive のみ容認
　　［2］to-infinitive と of + ing 共に容認
　　［3］of + ing のみ容認
（Quirk et al., 1985, pp. 1272-1274；安藤，2007, pp. 782–784）

（5）He *had the advantage* (over me) *of* knowing the language.
（G 大，2001，“advantage”）

名詞 advantage が従える同格を表す後置修飾のパタンを the Corpus of Contemporary American English（以後，COCA; Davies, 2008-）と the British National Corpus（以後，BNC; Davies, 2004）で調べてみると，以下の表 1 に見られるように上述（4）の［2］のパタンに該当し，特に HAVE the advantage of Ving の形が両コーパスから容易に検索される。[2] そして，（6）の実例からも of 以下が advantage（強み・長所）の内容を表しており，of が「同格」として機能しているといえる。尚，(6a）の文は，先に示したイギリス人ネイティブスピーカーが否定的に捉えている（2a）と同じ形式の実例である。

表 1. 名詞 advantage が従える同格を表す後置修飾の実態

	of Ving	to V
COCA	961 件	7 件
BNC	263 件	0 件

（6）a. … , he will *have the advantage of* being elected as a candidate of reform.
(COCA, Magazines)
b. A telephone call *has the advantage of* giving a personalised response, and yet is relatively inexpensive and not time consuming. (BNC, Written)

2.2　advantage が従える同格 that

一方, advantage が同格 that を従えるという記述は, 綿貫･ピーターセン（2006, p. 234）が「同格の that」を従える名詞として advantage をリストアップしているが，辞典や参考文献で該当すると考えられる用例をあげているのは（7）に見られる程度である。

（7）He *had the advantage* over me *that* he could speak English.
（ランダムハウス，1973，“advantage”）

先行研究が少ないため，ここでは統語的振る舞いから advantage が同格用法の that を従える可能性を探る。2.1 節で触れたように advantage が同格 of を従えるという事実がある。一般に，同格節を従える名詞は（8）のように BE 動詞で結ばれる論理関係にあるといわれる。実際，advantage もコーパスで同形式の用例（(9) 参照）が検索される（COCA では 71 件，BNC では 11 件）ことから，同格の that 節を従える可能性はあるといえる。事実，（9a）の advantage（長所）は，必要な素材の量がとても少ないことであり，（9b）における強みは，たった 1 本の光ファイバーで適切なライトを作動するための全情報を伝達することができる，ということで，それぞれ BE 動詞が従える that 節が長所や強みの内容を説明している。

（8）The *possibility* is that he will resign as ambassador.　（安井，1996，p. 72）
（9）a. There are many ways of making thin film cells; the *advantage* is that the amount of material needed is very small and that reduces the cost.
（COCA, Blog）
b. The *advantage* is that just one fibre can convey all the information to operate the correct lights.　（BNC, Written）

2.3　HAVE the advantage that 節内における助動詞 CAN の必要性

2.2 節で advantage が同格の that 節を従える可能性について触れた。この同格を表す that 節内に助動詞 CAN が必要かどうかという先行研究はないが，辞典には（10）のような could を含む用例が見られる。また，（11）の WordbanksOnline（以後，小学館 WB・WB; Harper Collins Publishers in Shogakukan Corpus Network, n.d.; Harper Collins Publishers, n.d.）の検索例を確認された

い。[3] コーパスでは成句 HAVE the advantage of に続く形として，CAN と類義の「内在的能力」「状況的能力」を表す being able to V を検索するのは難しくなく，“能力”という意味要素を従える可能性が観察される。[4] 第 3 節以降で詳細に分析するが，確かに（12）のように同格の that 節内に can が共起する実例が見られる。

（10）He *had the advantage* over me *that* he could speak English. （=（7））

（11）He *has the advantage of* being able to play in several forward roles.
（小学館 WB, UK, Written）

（12）They（= Acrylic varnishes）are very quick drying, with a low odour, and *have the advantage that* the brushes can be cleaned with water. （BNC, Written）

先行研究と上述の実態を踏まえ，次節から成句 HAVE the advantage of と比較しながら，以下 2 点の実態について実証的に調査していく。

（13）a. HAVE the advantage が 同格の that 節を従える可能性
b. 従える同格 that 節内の助動詞 CAN の必要性とその有無の要因

3. 量的調査

本節では，HAVE the advantage の後にどのような形式が来るのか，また同格 that 節がどのような頻度で生起するのか，その実態を数値から調査する。

3.1 HAVE the advantage of Ving の V の種類

成句として HAVE the advantage に同格 of が後続する形が無標といえるため，同格 of に続く動詞の種類を確認し，“能力”の意味につながる表現が生起するのか頻度面から調査する。

コーパスでは，（14）から（17）の各実例の下線に含まれるように of に後続する動詞の種類は多岐に渡り，being が最も多く，完了の having，状態動詞の knowing，動作動詞 allowing や reading などの頻度が高いといえる。

（14）a. Dana Perino *has the advantage of* being an attractive female.
（COCA, Blog）

b. Made of carbon, these cells *have the advantage of* being produced at low cost ··· . （COCA, Newspapers）

c. They *have the advantage of* being able to gain the trust of other residents, ··· . （COCA, Newspapers）

（15）I also *had the advantage of* having gone to an experimental progressive school ··· . （COCA, TV/Movies）

（16）··· George W. Bush will *have the advantage of* knowing these lessons from the past. （COCA, Magazines）

（17）a. Training of this type *has the advantage of* allowing us to share knowledge with a group of people at once, ··· . （COCA, Academic）

b. My Lords, I have *had the advantage of* reading in draft the speech prepared by my noble and learned friend, Lord Bridge of Harwich. （BNC, Written）

以下の表1-1と表1-2のCOCA，BNC，WBの調査から，共通してbeingを従える用例が他の動詞を含む用例に比べ圧倒的に多いことが分かる。そのbeingに続く形式を精査してみると，最も頻度が高いのが（14a）のような名詞句や形容詞が後続するパタンであり，次いで（14b）の受動態と本稿で探っている「内在的能力」「状況的能力」を表す（14c）のable to Vで，able to Vの出現率は全体で約20％であることが分かった。また，レジスターの相違から比較すると，HAVE the advantage ofが書き言葉の表現であること，英米差は見られないことが明らかとなった。

表1-1. COCA及びBNCにおけるHAVE the advantage of VingのVの種類

	COCA				BNC			
順位	件数	Vの種類	being + α		件数	Vの種類	being + α	
1	246 (1)	**being**	名 / 形	150	88 (1)	**being**	名 / 形	54 (1)
2	44	**having**	受動	37	34	reading	受動	16
3	25	**allowing**	**able to**	**36(1)**	9	**providing**	**able to**	**12**
4	19	**knowing**	存在	23	9	making	存在	6
5	11	**providing**			8	**having**		

注：太字は他コーパスでも検索された動詞を表示。

：表中( )の数値は話し言葉の件数。

：網掛けは本稿でターゲットにしている意味表現（able to V)。表1-2も同様。

表 1-2. WB における HAVE the advantage of Ving の V の種類

順位	件数	V の種類	US：being＋α		件数	V の種類	UK：being＋α	
1	36(2)	**being**	名 / 形	25(2)	99(7)	**being**	名 / 形	57(6)
2	8	**having**	able to	**7**	20	**having**	able to	**21**
3	4	**knowing**	受動	4	9	**knowing**	受動	15(1)
4	3	**allowing**			5	playing	存在	6

また，being able to に後続する動詞の種類は，以下の表 2 に示す動作動詞となるが，そのバリエーションは多岐に渡る。その一例が，（18）の include を含む実例である。

表 2. being able to に後続する動詞の種類

順位	COCA：36 件		BNC：12 件		WB の US と[UK]：28 件	
1	**use**	2(1)	**use**	1	[tap]	2
2	**operate**	2	**operate**	1	[work]	1
3	**look (at)**	2	offer	1	examine	1
4	jump-(start)	2	**include**	1	[**look (at)**]	1
5	**include**	1	snatch	1	[**use**]	1

注：各コーパスにおける総件数と表中の個々の動詞の合計件数が一致しないのは，各コーパスの上位 5 つの動詞のみを掲載しているため。

：太字は他コーパスでも検索された動詞，(　) の数は話し言葉の件数を表示。

：WB の欄に見られる[　]はイギリス英語から検索された動詞を表示。

（18）These experiments *have the advantage of* being able to include service attributes that have not been offered in real-world markets, ...（COCA, Academic）

3.2 HAVE the advantage that S' +（CAN）+ V' の可能性

既に 2.2 節と 2.3 節で概観したように，先行研究に名詞 advantage が同格の that 節を従える可能性と CAN の必要性を記述したものは極めて少ない。しかしながら，各コーパスからは同格の that 節を従える実例や同格の that 節内に CAN が生起する実例が検索される。（19）では，地元の組織ならではの協力体制がすぐに生まれるという強みを that の同格節で説明している。興味深いのは，話し言葉のデータである用例（20）である。一度，has the advantage of と言い，of を that 節に言い換えていることからも，前置詞 of とこの that 節が共に「同格」

として advantage の内容を説明する機能を果たしていることの例証になるといえる。（21）は，同格を表す that 節内に can を従える用例で，各用例は風力発電の 24 時間稼働し続ける“能力”の高さや ID が複数の銀行で利用“可能”な長所を表している。

（19）Local organizations *had the advantage that* cooperation came about quickly because people knew and trusted one another. （COCA, Academic）

（20）Making a will with a solicitor also *has the advantage of that* you'll have a copy, he's likely to look after it for you if you want him to. （BNC, Spoken）

（21）a. Wind generators also *have the advantage that* they can go on working 24 hours a day, where solar panels can only operate in sunlight.
（BNC, Written）

b. ID card *has the advantage that* it can be used with more than one bank.
（小学館 WB, UK, Written）

数値面を詳しく見てみると，以下の表 3 が示すように COCA では，同格の that 節を従える 83 件の内 20 件が CAN と共起し（CAN を従える割合は約 25％），BNC では 95 件の内 16 件が CAN をとる形式（約 17％）であった。WB のイギリス英語では，同格を表す that 節を従える 57 件中 18 件（約 30%）で CAN と共起し，その内の 7 件が話し言葉であった。WB のアメリカ英語では，19 件の同格節のうち 5 件（約 20%）が CAN を節内にとる用例であった。3 つのコーパスで検索された CAN が生起する件数からは，イギリス英語で使用される傾向があるように見えるが，英米差はほとんど見られない。実際，石川（2008, pp. 83–97）を参考にカイ二乗検定にかけると英米差で有意差は確認されなかった。[5] コーパスのデータでは英米差に関して有意差が見られなかったため，第 5 節のインフォーマントの反応も見ながら再確認していくことにする。英米間での明らかな使用頻度差は見られなかったが，HAVE the advantage が同格の that 節を従えること，その節内に CAN を従える形式は英米問わず書き言葉で主に利用される傾向にあること，イギリス英語では話し言葉でも使用されることが明らかになったことは注目すべき点である。

表 3. HAVE the advantage that S’ +（CAN）+ V’ の実態

	COCA	BNC	WB [UK]	WB [US]
同格 that 節	83	95 (2)	57 (14)	19
＋ CAN	20	16 (1)	18 (7)	5

注：[UK][US]はイギリス英語とアメリカ英語，( )の数は話し言葉の件数を表示。

更に，“能力”を表す CAN と共起する動詞を精査してみると，表 4 が示す通り，受動や能動といった態の違いに関わらず，先に見た being able to と同様に動作動詞であることが分かる。

表 4. COCA：HAVE + the advantage that S’ + CAN + V’ における V’ の種類

受動 be +	seen	used	not called	represented	estimated	changed	
能動	adapt	make	stimulate	fabricate	summarize	examine	identify
	provide	evaluate	take out	respond to	get out		

3.3 量的調査のまとめ

ここまでの量的調査から，以下 3 点を実証することができたと考える。1 つ目（(22a)）は，両表現とも動作動詞を従え書き言葉で利用される傾向にあるということである。2 つ目（(22b)）に，本研究がターゲットとする HAVE the advantage of being able to V / that S’ + CAN + V’ の両形式の出現率に英米差はなく，イギリス英語では話し言葉でも利用される傾向があるということである。3 つ目（(22c)）として，同格 that を従える節の内，“能力”を表す CAN が生起する割合は約 25％に過ぎず，必ずしも advantage（強み・長所）にとって必要不可欠な意味要素ではないということである。

（22）a. HAVE the advantage of being able to V / that S’ + CAN + V’ の両表現とも動作動詞を従え書き言葉で利用傾向

b. HAVE the advantage of being able to V / that S’ + CAN + V’ の両表現の頻度に英米差はなく，イギリス英語では話し言葉でも利用傾向

c. HAVE the advantage が同格の that 節を従える実例の内，that S’ + CAN + V’ のパタンの出現率は約 25％で，助動詞 CAN は advantage にとって任意要素

確かに，先行研究では触れられていない HAVE the advantage が従える同格 that の節内に“能力”を表す助動詞 CAN が生じることが実例から明らかとなり，本稿の冒頭で触れたイギリス人ネイティブスピーカーの指摘が正しい側面もある。しかしながら，同格の that 節内における CAN の出現率を確認すると，必須要素とは言えず，ネイティブスピーカーの直観と言語実態が必ずしも一致していないことも分かる。では，どのような要因が同格の that 節内に CAN を生起させるのか，また比較対象の類義表現 being able to V がなぜ生じるのか，について次節で実例を 1 つ 1 つ検討しながら論じていく。

4. 質的調査

4.1 HAVE the advantage that S’ +（CAN）+ V’ の生起条件

以下（23）は, HAVE the advantage that S’ + V’ の形式の実例である。(23a）は，波線部の often contain（よく～を含んでいる）という表現からも習慣かつ状態を表すのは明らかで“能力”の意味とは共起しないと考える。(23b）では，同格節内の動詞の形から Herman の常日頃の話し方について述べており，(23c）も外付けのドライブであれば普通はコンピューターの故障や不慮な事故で影響は受けない，と共に習慣や状態を表す内容であることから，“能力”を表す CAN を特に表現する必要はない。(24) はどうであろうか。(24) の命題内容は，波線部 came about が「(予想外のこと）が起こる・生じる」と制御できない事象の発生を表し，能力とは関係ないことに加え，地元の人々がお互いを知り信頼し合っているため習慣的に協力体制が生まれていると考えることができる。いずれにせよ，能力と無関係であり習慣とも読み取れることから，(23）の各例と同様に CAN を用いる場面ではないと説明できる。

（23）a. Old photos and videos *have the advantage that* they often contain little gems that you had entirely forgotten about.（COCA, Web）

b. HUCKABEE: Herman Cain does *have the advantage that* he talks the language of people who understand that he’s talking to them and he’s talking for them.
O'REILLY: He’s a good communicator.（COCA, Spoken）

c. External drives also *have the advantage that* they usually are not affected by computer crashes and other mishaps; … .（COCA, Web）

（24）Local organizations *had the advantage that* cooperation came about quickly because people knew and trusted one another. （=（19））

一方, 次に示す（25）の用例は, HAVE the advantage that S' + CAN + V' の形式, つまり“能力”を要求する実例である。（25a）は, 主題の風力発電が, 波線部で示す日が当たっている時間帯でのみ機能する太陽光パネルと比較することで, 1 日 24 時間稼働し続ける能力という強みについて述べられている。（25b）も, 人間と比べて機械には新しい環境に即座に適応することができるという長所を示している。この 2 つの用例には, 共通して主語が元来保有する“内在的能力”の意味が他の事象と比較され焦点化されており, それを can で表現しているといえる。

（25）a. Wind generators also *have the advantage that* they can go on working 24 hours a day, where solar panels can only operate in sunlight. （=（21a））

b. Machines do *have the advantage that* they can within a geologic instant adapt to a new environment. （COCA, Web）

では, 以下（26）に含まれる助動詞 can の意味はどうか。（26a）の内容は, アクリルニスの特性を説明したもので, アクリルニスであればブラシの汚れが水できれいになるという長所がある, という。（26b）は, ソフトウェアモデルを利用すれば必要なほどモデルを複雑にすることができるというソフトウェアモデルの長所を述べている。先の（25）の用例と異なり, （26）の 2 例は, 長所を持つ主語と that 節内の命題の主語が異なることからも, 内在的能力ではなく, 強みや長所を有している事物の状況的能力を表していると考える。

（26）a. They（= Acrylic varnishes）are very quick drying, with a low odour, and *have the advantage that* the brushes can be cleaned with water. （=（12））

b. Software models *have the advantage that* you can make the models as complex as you need to, ⋯ . （COCA, Blog）

上述のように, HAVE the advantage that に後続する節内に助動詞 CAN が生起するかどうかは, 名詞 advantage の意味が引き出しているのではなく, むしろ節内の命題内容に起因していると考える。つまり, 同格節内の主語が他のもの

と比較して能力を明示する場合や強調する場合，また状況的により能力が発揮されることを示す場合にCANを必要とするのである。

4.2 HAVE the advantage of Vingかbeing able to Vの選択

類似した意味で利用されるHAVE the advantage of being able to Vの生起条件が何なのか確認してみる。（27）の各用例には同じ動詞playが用いられている。（27a）の例は, HAVE the advantage of Vingの形式で, 将来有望なバスケットボールのPrice選手が強力なポイントガードであるBrown選手と常日頃から一緒にプレーしている利点を示している。一方，HAVE the advantage of being able to Vの形式で表現されている（27b）は，フットボール選手である主語Heが他の選手と比べ, フォワードのポジションにおいて様々な役割（ポストプレーなど）でプレーできるという強みを持っている，という内容である。この2例に見られる大きな違いは，習慣的な行為なのか他者と比べて能力の違いを強調しているかという点で，前者の意味では“能力”を必要とせず，後者の意味ではbeing able toで表現されるのがより自然といえる。これは，4.1節で論じたHAVE the advantageが従える同格節that内の助動詞CANの有無と同じ規則で説明できると考える。

（27）a. Price has a bright future as a player and *has the advantage of* playing with Brown, a strong point guard. （COCA, Newspapers）

b. … he is very professional, a charming man and a good guy to have as a team-mate. He *has the advantage of* being able to play in several forward roles. （小学館WB, UK, Written）

もう少し実例を探ってみる。（28a）はロシア語を話す強みについて述べている。この用例では一見,「ロシア語を話すことができるという強み」と解釈でき，being able to speak Russianとした方が良さそうだが，Swan（2005, p. 106）が示す通り, 動詞speakやplayは助動詞CANなしでも能力を表すことができるため，ここではbeing able toを特に明記する必要はない。（28b）では，下線が示す動詞knowが状態動詞のため，能力を表すcanとは共起せず，（28c）では，「視覚に入ってくる」という無意識的な知覚を表す動詞seeが用いられており，（28b）と同じ説明が適用できる。ダーウィンには鳥たちが何を食べていたのか目にしたといった利点が（当時は）あったが，包括した進化の概念は化石記録

の研究に基づいている，という解釈となり，see が状態的な動詞の役割を果たしていることから，“能力”の意味とは結びつかないと考える。

（28）a. For instance, if one says to them, you at least *have the advantage of* speaking Russian and this entitles you to a certain authority in the area of Eastern European art, … . （BNC, Written）

b. … George W. Bush will *have the advantage of* knowing these lessons from the past. （=（16））

c. Darwin *had the advantage of* seeing what the birds ate, but the whole evolution concept is grounded in the study of fossil records. （COCA, Blog）

4.3 質的調査のまとめ

本節では，コーパスの実例から，主に HAVE the advantage that S’ +（CAN）+ V’ のパタンに焦点を当てて，成句表現 HAVE the advantage of Ving / being able to V とも比較分析しながら助動詞 CAN と共起する要因を探ってきた。その分析結果は，（29）の 3 点に集約することができる。助動詞 CAN そして being able to V が HAVE the advantage that / of に続く要因の根底には，命題中に他の事象と相対的な主題の能力（内在的能力）の明示・強調や能力が発揮される状況（状況的能力）といった要素が存在している。逆に，習慣的な行為や状態を表す動詞では，助動詞 CAN や being able to V といった表現と共起しないのは至極当然だといえる。

（29）a. HAVE the advantage that S’ +（CAN）+ V’ と HAVE the advantage of Ving / being able to V のパタンの選択は共に同じルールが適用可能

b. 他の事象との比較による能力の明示・強調や状況的能力の発揮を表現する場合に CAN や being able to と共起傾向

c. 習慣的な行為・状態動詞（相当）の場合には助動詞 CAN 及び being able to と共起しない傾向

5. インフォーマント調査

5.1 調査方法の概要と予測

これまでの量的・質的調査結果の妥当性を確認するため，インフォーマント

調査（アメリカ英語：アメリカ人3名とカナダ人1名・イギリス英語：イギリス人6名の計10名）を行った。(30）から（32）の用例は，コーパスを参考に著者が加工して作成した質問文である。各質問文の後ろには，これまでの調査結果に基づいた著者の予測が示され，調査結果は表5に示されている。調査方法は，各質問文に対して，被検者であるネイティブスピーカーが，容認できる場合は○，容認できない場合には×，そして，容認できないわけではないが不自然さを感じる場合に△，という3つの記号（選択肢）で回答した。

各質問文には調査から得られた意図が含まれており，著者の予測はそれに基づいてなされている。例えば，(30）の質問文は，冒頭のCompared with face-to-face classesという他の授業を比較してonline classesの長所を強調している点でbeing able to Vと同格のthat節内に助動詞CANが共起している文の容認度が上がると推測される。また，(30a）と（30c）における予測の相違は量的調査の頻度（Ving / beingとthat節の件数）に基づいている。一方，(31）の質問文にはcontainという状態動詞を利用している点で“能力”を表す表現とは共起しないため，(30）とは正反対の結果，つまりbegin able toとcanが含まれていない文の容認度が上ると予測した。そして，(32）は理由を明示するbecause節を伴うことで，海外勤務の能力を有するという強みを表しているため，begin able to及びcanを用いた文の容認度が上ることを想定した。

(30) a. Compared with face-to-face classes, online classes have the advantage of being taken at any time.　著者の予測：△

b. Compared with face-to-face classes, online classes have the advantage of being able to be taken at any time.　著者の予測：○

c. Compared with face-to-face classes, online classes have the advantage that they are taken at any time.　著者の予測：×

d. Compared with face-to-face classes, online classes have the advantage that they can be taken at any time.　著者の予測：○

(31) a. Old albums have the advantage of containing little gems that you had entirely forgotten about.　著者の予測：○

b. Old albums have the advantage of being able to contain little gems that you had entirely forgotten about.　著者の予測：×

c. Old albums have the advantage that they often contain little gems that you had entirely forgotten about.　著者の予測：○

d. Old albums have the advantage that they can often contain little gems that you had entirely forgotten about. 著者の予測：×

（32）a. They have the advantage of working overseas because they speak English fluently. 著者の予測：△

b. They have the advantage of being able to work overseas because they speak English fluently. 著者の予測：○

c. They have the advantage that they work overseas because they speak English fluently. 著者の予測：×

d. They have the advantage that they can work overseas because they speak English fluently. 著者の予測：○

5.2 回答結果

各質問文を参照しながら，回答結果と著者の予測について分析していく。まず，比較対象を示して online classes の長所を強調している（30）の各質問文では，被験者の回答は著者の予測から大きく外れるものではなかったが，HAVE the advantage of に続く Ving か being able to V かの容認度は質的調査結果と一致しなかった。予測に反し（30a）の HAVE the advantage of being taken の形式における容認度が高く，対面授業と比較はしているものの（30b）の HAVE the advantage of being able to be taken を不自然とする回答が多かった。[6] ネイティブスピーカーのコメントには，begin able to be taken の表現の長さや時制の複雑さに不自然さを感じるとの指摘が見られた。興味深いことに，イギリス英語では HAVE the advantage of に Ving が後続する形式は容認されるのに対し，その言い換えとも言える同格の that 節を従える (30c) の HAVE the advantage that S’ + V’ の形式では 6 人中 5 人が非容認と回答した。同様に，HAVE the advantage that S’ + CAN + V’ の形式は容認傾向にあるのに対して，HAVE the advantage of being able to V では不自然とする回答が増えており，インフォーマントによる指摘の通り，意味だけでなく構造などの問題点も考慮する必要があるというインフォーマント調査の今後の課題と言える。

状態動詞を利用した (31), そして理由を明示した (32) の質問文へのネイティブスピーカーの反応は，これまでの調査結果から著者が予測したものとほぼ一致するものであった。状態動詞 contain を含む質問文（31）では，“能力”を表す表現を含む（31b）及び（31d）で容認度が低かった。一方，（31a）と（31c）の“能力”を示さない文は容認傾向にあるが，僅かながら同格の that 節を従え

る用法に不自然さを感じる傾向が英米共に垣間見える結果となった。（32）の質問文では，理由を明記することにより海外で仕事ができるという“能力”の高さを際立たせており，being able to と can を含む英文の容認度が明らかに高いといえる。特に，（32c）と（32d）の同格節を従える両質問文における容認度の差は顕著であり，意味と形式の選択が密接に結びついていることが分かる。

表 5. インフォーマント調査の結果

設問		a	b	c	d
（30）	米	2-1-1	1-3-0	0-2-2	3-0-1
	英	4-2-0	2-3-1	1-0-5	4-0-2
（31）	米	3-1-0	0-1-3	2-1-1	1-1-2
	英	4-0-2	0-4-2	3-1-2	2-1-3
（32）	米	1-2-1	4-0-0	0-3-1	3-1-0
	英	4-1-1	5-0-1	0-0-6	4-2-0

注：表中の数値は，容認[○]—不自然[△]—非容認[×]，の順で提示。

：「米」「英」はそれぞれ「アメリカ英語」「イギリス英語」のインフォーマントを表示。

5.3 インフォーマント調査のまとめ

本節では，インフォーマントの反応を調査することにより，第 3 節の量的調査と第 4 節の質的調査で分析した結果の妥当性を確認し，以下（33）に示す 3 点にまとめることができると考える。1 つ目は，文構造や時制の複雑さとの関係について更なる調査が必要だが，（30）の設問以外では，英米共に HAVE the advantage of の形式より同格の that 節を従える形式自体，容認度が下がる傾向にあり，量的調査の傾向に類似した結果であった。2 つ目は，名詞 advantage が必ずしも“能力”の意味を引き出しているわけではなく，むしろ対象物と比較することで主題が行う“能力”の高さを明示・強調する場合や理由等を提示し，ある状況下で能力が発揮される場合に，being able to V や同格 that 節内に CAN を用いる傾向にあることが再確認できた。3 つ目は，状態動詞や習慣を表す場面（動詞）では，“能力”を表す表現とは共起しないという傾向がはっきりと表れ，（33b）同様，質的な分析結果の妥当性を裏付けるものとなった。

（33）a. 英米共に HAVE the advantage of より同格の that 節を従える形式で容認度が下がる傾向があり，量的調査の結果と類似

b. 他の事象と比較すること，また命題内容に対する理由を提示するこ

とで，“能力”の高さや状況的能力を明示する場合，HAVE the advantage が of being able to V や that S’ + CAN + V’ の形式と共起傾向

c. 状態動詞や習慣的な長所や強みを表す場合，HAVE the advantage に続く表現は Ving や that S’ + V’ の形式で，“能力”の表現を明示すると容認度が下がる傾向

6. 結語

本稿では，コーパスをデータの中心に据えた実証的なアプローチで，既存の成句 HAVE the advantage of Ving と比較しながら，これまで扱われることのなかった名詞 advantage が同格の that 節を従えることが可能なのか，更に同格節内に“能力”を表す助動詞 CAN が必須要素なのか論じてきた。結論として，HAVE the advantage that S’ + CAN + V’ の形式は英米共に主に書き言葉で利用されるが，イギリス英語では話し言葉でも利用され，アメリカ英語と比べより広いレジスターで利用されている傾向にあることが明らかになり，本稿冒頭で触れたイギリス人ネイティブスピーカーから指摘があったことも合点がいく。しかしながら，同格の that 節内に生起する助動詞 CAN は，決して advantage の意味が引き出しているわけではなく，of や that が従える命題内容に依存していることを実例を見ながら実証した。そして，インフォーマント調査でも分析結果の妥当性がある程度認められたように，“内在的能力”“状況的能力”を明示・強調する際に，“能力”を表現する CAN や類義の being able to が生起するというメカニズムが根底にあることが明らかとなった。つまり，助動詞 CAN は advantage の必須要素ではなく，同格節内の命題内容に依存しているのである。

本稿で明らかにした HAVE the advantage that S’ +（CAN）+ V’ という新たな知見（表現）は, これまで文献や辞典にも詳細な記述がなされていないことから, 言語実態をより正確に反映した記述を辞典に提供できる点で貢献できる。今後, 本研究結果を基盤に，merit，benefit，strong point などの advantage の類語も同じパタンを従えることが可能なのか，更に調査と分析を進め，意味と形式の関係性やそこに働く原理を更に探っていきたいと考える。

注

* 本稿の執筆に際し，特に形式面と統計面で大変貴重なご指摘とご助言を賜りました

査読委員の先生方に心より感謝申し上げます。また，2024年10月5日に青山学院大学で開催された第50回大会の口頭発表において，フロアから貴重なご意見を下さった京都外国語大学名誉教授の赤野一郎先生にも感謝の意を記します。尚，未だに残る不備な点は全て著者の責任である。

1. 本稿でHAVE the advantage of ... のHAVEを大文字で表記しているのは，人称と数の一致や過去時制によるhave，has，hadを全て含むことを表している。また，CANもcan及びcouldを含むことを示し，BE動詞も同様である。

2. Advantage（長所・強み）の内容を説明する同格の後置修飾のパタンには，コーパスによる検索数は非常に少ないが，HAVE the advantage to Vというto Vの形式も存在する。この形式も先行研究では触れられておらず，HAVE the advantage of VingとHAVE the advantage to Vの意味の違いについても興味深いので別の機会に調査したいと考える。

3. 本稿の調査では, 厳密に言うと2種類のインタフェースでWBを利用した。1つは, 2024年3月まで提供されていた小学館コーパスネットワークを介して利用したWB，もう1つは2025年2月から利用を開始したHarper Collins Publishersが提供しているオンラインプラットフォームのSketch Engine内で検索したWBである。検索結果から，両者は収録語数だけでなく収録されている用例も異なるため，本稿では出典として前者を“小学館WB”，後者を“WB”という形で区別した。本稿で提示している用例の多くは前者から，表1-2，表2及び表3のデータは後者からのものである。

4. 本稿で述べる“内在的能力”と“状況的能力”のCANとは, 一般的に“能力 (ability)”と“状況的可能性（circumstantial possibility）”と定義されるもので，以下の相違が見られる：

［1］I can speak a little Italian. 能力　（安藤，2005，p. 275）

［2］You can speak English tomorrow / here.　状況的可能性（本稿の状況的能力）　（安藤，2005，p. 278NB1）

［3］I can swim, but today I can’t because the sea is rough.　(G6, 2022, “can”)

上記［3］の用例では，前半が“内在的能力”，後半が“状況的可能性”を表すCANの用法である。類義のBE able toにも上記の両用法が同様に存在する。

尚, 研究当初, 本稿で議論してきた助動詞CANの意味を先行研究に基づき“能力”と“可能”（状況的可能性）として捉えていた。しかし，類義表現であるBE able toの形容詞ableの原義との整合性を図り，後者の意味が，“ある状況下”においてadvantage（強み・長所）を持つ“事象の能力”が発揮されるかどうかで多くの用例を捉えることが可能なため，本稿では“状況的能力”の用語を用いて論じることにする。

5. 表3の「HAVE the advantage that S’ +（CAN）+ V’ の実態」に関して，小学館コーパスネットワークが提供する小学館WBが2024年3月末でシステムの提供が終了し，

審査段階で検索ができていなかったため，WB［UK］と［US］の同格 that 節の件数が“未確認”と表記されていた。しかしながら，英米差を含めより正確なデータを掲載できるよう Harper Collins Publishers が提供している WB で調査してそのデータを追加して修正した。また，査読委員の先生より，COCA と BNC における同格の that 節内に CAN が生起する形式に関して，カイ二乗検定にかけると英米差に有意差は確認できないとのコメントを頂いた。そのため，WB を新たに含めた 3 つのコーパスで上述の形式における英米差をカイ二乗検定にかけて再調査した。より正確な実態を提示する貴重なきっかけを与えて下さった査読委員の先生方に感謝申し上げます。

6. (30a)は, 本稿の冒頭でイギリス人ネイティブスピーカーに指摘を受けた用例((2a))と同じ構造の文である。4 名のイギリス人が容認している通り，個人差も影響するインフォーマント調査の難しさを感じる。文構造を含め，今後，更に調査を継続していく必要がある。

参考文献

安藤貞雄.（2007）. 現代英文法講義. 開拓社.

Biber, D., Johansson. S., Leech. G., Conrad. S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman.

Davies, M. (2004). *The British National Corpus* (BNC) (from Oxford University Press). Available online at https://www.english-corpora.org/bnc/ 2023 年 6 月～2025 年 2 月アクセス

Davies, M. (2008-). *The Corpus of Contemporary American English* (COCA). Available online at https://www.english-corpora.org/coca/ 2023 年 6 月～2025 年 2 月アクセス

Haper Collins Publishers. (n.d.). *WordbanksOnline* (WB). Available online at https://wordbanks.harpercollins.co.uk/ 2025 年 2 月アクセス

Haper Collins Publishers. (n.d.). 小学館 *WordbanksOnline*（小学館 WB）. In Shogakukan Corpus Network. https://scnweb.japanknowledge.com/ 2023 年 6 月～2024 年 3 月アクセス

井上永幸・赤野一郎（編）.（2018）. ウィズダム英和辞典（第 4 版）. 三省堂.（W4）

石川慎一郎.（2008）. 英語コーパスと言語教育. 大修館書店 .

小西友七・南出康世（編）.（2001）. ジーニアス英和大辞典. 大修館書店.（G 大）

Lea, D., & Bradbery, J. (Eds.). (2020). *Oxford Advanced Learner's Dictionary* (10th ed.). Oxford University Press. (OALD10)

南出康世･中邑光男（編）.（2022）. ジーニアス英和辞典（第 6 版）. 大修館書店.（G6）

Quirk, R., Greenbaum. S., Leech. G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

Rundell, M (Ed.). (2007). *Macmillan English Dictionary* (2nd ed.). Macmillan Publishing. (MED2)

小学館（編）.（1973）. ランダムハウス英和大辞典. 小学館.（ランダムハウス）

Swan, M. (2005). *Practical English Usage*. Oxford University Press.

綿貫陽・マークピーターセン．（2006）．実践ロイヤル英文法．旺文社.

安井稔．（1996）．コンサイス英文法辞典．三省堂 .

（土屋　知洋　芝浦工業大学　E-mail: tomotsu@shibaura-it.ac.jp）

「論文」

“Dictions” *by Two Brothers*, Charles and Alfred Tennyson

Iku FUJITA

Abstract

This study employs a quantitative method, latent Dirichlet allocation topic model, to examine the distinctive thematic and lexical characteristics of poems by Alfred, Lord Tennyson, and his brother, Charles Tennyson, focusing primarily on *Poems by Two Brothers* (1827), the first published collection by Alfred, Lord Tennyson. It has been said that there is some ambiguity surrounding the poems’ authorship within the collection, originally published without annotations indicating each poem’s author. This research leverages topic model to uncover patterns in the diction and thematic inclinations of the two brothers. The results of the latent Dirichlet allocation analysis indicate that each poet gravitated toward certain specific topics as dominant themes in his works. Topic 17, which is associated with romantic sentiments and primarily physical descriptions of women, is prevalent in the poems written by Charles Tennyson, while Topic 1, which concerns themes of masculinity, enthusiasm, and battle, is prominent in the poems written by Alfred Tennyson. These findings highlight the distinctive differences between the two brothers in word usage. It is noteworthy that this study represents a novel application of topic model in examining characteristic diction of the two poets, offering internal evidence of the distinct word usages within the Tennysons’ collaborative collection. This exploration underscores that topic model is effective in distinguishing thematic tendencies as well as the characteristic diction of the two authors, Alfred and Charles Tennyson.

1. Introduction

This study investigates the characteristic content of two poets, Alfred, Lord Tennyson (1809–1892) and his brother, Charles Tennyson Turner (1808–1879), using a quantitative method, the latent Dirichlet allocation (LDA) topic model (Blei et al., 2003). *Poems by Two Brothers* (1827)[1] is the first collection of 19th-century poet, Alfred Tennyson and his brother, Charles Tennyson. The authors of the poems in the collection had been considered two, Alfred and his brother, Charles, when *Poems by Two Brothers* was published. There was, however, no signature or annotation that referred to the author of each poem. The second edition of the collection was published in 1893, soon after the death of Alfred. The most significant change from the first to the second edition was the capital letters at the end of each poem suggesting its author. These notations clarify not only which poems were written by Charles or Alfred but also suggest the cooperation of another author, Frederick; interestingly, some works remain unidentified.

The authorship attribution of the *Poems by Two Brothers* as well as the collection itself have attracted the attention of few scholars. Brimley (1972) mentions that Alfred is responsible for roughly half of the poems in the collection. Paden (1964) qualitatively challenges the matter of unidentified authors by comparing capital letter annotations and notes written in the two manuscripts (Haddelsey's and Charles's copy) and the second edition of *Poems by Two Brothers*. Paden (1964) further identifies the authorship of 18 poems out of 20 poems, whose authorship are marked as uncertain or doubtful poems and for which the initials on two manuscripts and the 1893 edition do not match. Although there still remains two poems with uncertain authorship, it seems there is no study to follow Paden, perhaps because he had this to say about the collection as a whole: "None of the poems in question has any noticeable literary value, to be sure" (Paden, 1964, p. 147). This low estimation of the poems dovetails with criticism that the collection is largely imitations of fashionable styles at the time (Delphi Poets Series, *Alfred, Lord Tennyson*, 2013). In fact, successive studies on Alfred, Lord Tennyson do not fully address the authorship of poems in the collection, and it can be assumed that their perceived lower literary value might have lessened the interest of other scholars. Several collections of Alfred Tennyson's have been published, many of them edited by Christopher Ricks. The second edition of *The Poems of Tennyson* published in 1987 is compiled in the three volumes. Ricks lists six poems as "doubtful poems: poems attributed to Tennyson [Alfred]" in Appendix C of the third volume. Five of six poems, "Egypt," "The Deity," "On the Moon-Light Shining upon a Friend's Grave," "The Dying Christian," and "Switzerland," are marked as doubtful or uncertain works in the 1893 edition of the *Poems by Two Brothers* (Ricks, 1987, pp. 641–646). In the 1893 edition, the sixth poem, "Song [To Sit Beside a Christal Spring]," was assigned "A.T.," suggesting it is Alfred's work. This conclusion differs from that of Paden (1964), who claimed that the five poems except for "Egypt" are attributed to Charles.

Ricks states that authorship attribution should be more concerned with external evidence, such as annotation in manuscripts than with internal evidence, such as poetic diction (1987, p. 647). Nonetheless, depending solely on the external evidence is limited by the absence of reliability and/or of information in the collection's manuscripts. Paden (1964) considered the handwriting and annotations in the manuscripts, as well as the differences in style but found few explanations for specific stylistic elements that could be used to identify the authors of particular works.

Regarding stylistics, function words such as determiners, prepositions, and conjunctions are frequently highlighted. It is notable that quantitative authorship attributions as well as stylometry (quantitative stylistic analysis), often prioritize the examination of function words, or the most frequent words in a text, in their analysis of prose works. In contrast, content words (e.g., nouns, verbs, adjectives, and adverbs) receive comparatively less attention in authorship attribution studies. The rationale behind this is that the choice of content words is considered to reflect the content of the works, rather than the author's style. An analysis of poetic style will consider rhyme, meter, and rhythm. In particular, in rhyming (and other sound effects), syllable rhyme is constituted less by function words than by content words. Given this, it can be posited that the word usage and figurative expressions using content words, namely poetic diction, are related to the characteristics of poems by Alfred and Charles.

Several previous studies using quantitative approaches on the works of Alfred Tennyson's reveal that the 1820s poems use adverbs and nouns differently from other works written in the 1830s or later (Fujita, 2020, 2023). Fujita (2020) suggests a chronological difference in Alfred's use of *–ly* adverbs using Correspondence Analysis. Fujita (2023) utilizes LDA topic model to analyze Alfred's use of nouns. Fujita found that certain topics are concentrated in the 1820s poems. Both of Fujita's studies indicate that the authorship differences possibly caused the results, although Fujita (2023) did not separate Charles's and Alfred's works in her analysis.

The analysis of content words and/or function words, namely poetic diction, can be regarded as an examination of internal factors. This perspective opposes Ricks' suggestion. Nevertheless, if the analysis of this study can demonstrate discrepancies in internal elements through quantitative analyses, the results can help to augment the evidence found in previous studies and elucidate the divergences among authors. The use of a quantitative method engenders objective aspects and provides a divergent perspective from that of internal and external evidence. This study therefore employs a quantitative approach, the LDA topic model (Blei et al., 2003), which is adept in detecting semantic structures in the text data. This study aims to address two research questions: 1) Can LDA detect the differences in poetic diction between the works of Alfred and Charles? And 2) If LDA detects differences, what characteristics do the two authors exhibit? To investigate these questions, the study considers content words, which fill semantic roles and are intimately associated with poetic diction.

2. Data and methodology

2.1 Data

The target dataset (corpus) comprises 525 poems. The statistical description of the corpus is shown in Table 1. Among the 525 poems, 102 are from the first edition of *Poems by Two Brothers*, and 24 are unpublished works but are assumed to have been written in the 1820s by Alfred. The remaining 399 poems are Alfred's lyrical poems published/written in the 1830s to the 1890s. Each poem was assigned to a single file with filenames to indicate the author names, publication years, and abbreviated poem titles. The authorship of the poems, number of works for each category, and filename examples are shown in Table 2. The authorship of the poems from *Poems by Two Brothers* refers to the annotations in the second edition of *Poems by Two Brothers* (1893). The question mark ("?") with the initial of author's names (e.g., "A.T. (?)") suggests doubtful authorships. The initials of both Alfred and Charles (e.g., "A.T. or C.T.") indicates that it is doubtful which of Alfred or Charles is the author of the poems ("The Deity" and "The Dying Christian"). The 1893 edition proposed the existence of another author, Frederick, which the "A.T. or C.T." annotation further indicates was not the case, as the poems were not written by Frederick. "The Egypt" is the only poem with the note, "Begun C.T., finished A.T.," seemingly indicating that the work was begun and mostly written by Charles, with Alfred writing the last two stanzas (Paden, 1964).

The target corpus included the works from other collections and publication dates other than 1827. Analyzing poems not included in *Poems by Two Brothers* made it possible to ascertain whether any characteristics observed in Alfred's works were limited to a specific date or poems or if they were more pervasive throughout his career. While Alfred has a prolific oeuvre of over 50 narrative poems, this study excludes such works from the analytical purview. This is due to the fact that Tennyson's *Poems by Two Brothers* (1827) is composed exclusively of lyrical poems. The notable distinctions between Alfred's lyrical and narrative poems extend beyond the length of each work. They also encompass the themes, content, and characters present in the poems. It is therefore to be understood that narrative poems, while presented in the format of poetry, will exhibit content that is more closely aligned with that of prose. This discrepancy in word usage was noted by Fujita (2023). The subsequent analysis

Table 1. Descriptive statistics of the target corpus

Number of poems	525
The total tokens	168,282
Minimum number of words per poem	12
Maximum number of words per poem	18,662
Average number of words	320.54

Table 2. Breakdown of poems in the corpus and their filenames

	Authors[*1]	Number of works	Filenames
Poems in *Poems by Two brothers* (1827)	Alfred	46	A27_TITLE[*2]
	Alfred (?; doubtful)	4	a27_TITLE
	Charles	48	C27_TITLE
	Charles (?; doubtful)	1	c27_TITLE
	Alfred or Charles	2	ac27_TITLE
	Charles started and Alfred finished (mostly Charles)	1	CA27_TITLE
Poems in *Poems by Two brothers* (1893) but not in the 1827 edition	Alfred	3	Aup_title[*3]
Unpublished poems but thought to have been written in 1820s	Alfred	21	T20_title[*3*4]
Lyrical poems of Alfred Tennyson published in the 1830s–1890s	Alfred	399	e.g., T30_TITLE/title[*3*4]

*1 For the works from *Poems by Two Brothers*, the authors assigned in the 1893 edition are listed.

*2 Abbreviated titles of poems are inserted where TITLE suggests.

*3 The unpublished poems' abbreviated titles are indicated by lower case letters.

*4 The two digits following *T* indicate the last two digits of the publication year.

will therefore focus on Alfred Tennyson's lyrical poems to minimize the potential impact of genre on the results.

2.2 Methodology

The use of LDA allowed researchers to detect possible semantic links within the corpus under study and to categorize words that appeared in numerous documents into topics (Tabata, 2018, p. 52). The term *topic* here refers to a group of words, but it is not the same as the meaning used in field such as linguistics and literature. This method is considered to be well suited for content words, such as nouns that appear in a work's content and verbs that describe the movements of characters, because it uncovers latent semantic links among words based on their tendencies to co-occur. LDA employs the concept of the "bag of words," an approach that considers each document in the corpus as a bag (Jockers, 2014, p. 137). The larger the bag, the greater the likelihood of discovering words that are likely to co-occur within the same bag. The variance in document size is equal to that of bag size. With regard to the analysis of prose texts, Jockers posits that it is beneficial to divide novels (and other voluminous documents) into segments and then run the model (2014, p. 137). He asserts that LDA captures both overarching themes that traverse the entirety of a novel and themes that emerge and then recede at particular points.

Regarding segment size, namely the number of terms that compose a bag, Fujita (2022) proposed a relevant segment size for LDA practice in poetry analysis. This paper makes reference to Fujita (2022) to employ the mean token value (320 words, as illustrated in Table 1) of the corpus in question. The tokens in each poem are enumerated in sequence from the initial token to the concluding token, thereby segmenting each document. Once each text was divided into equal-sized consecutive segments, the two final parts were joined together, unless the final segment exceeded 160 words in length (which is the same as the half of the segment size). In the case of poems with a total number of tokens below 320, no duplicate segments were created; instead, each poem was treated as a single segment. Thus, the largest segment size was 480 words, while the smallest was 12 words. The 525 poems in the target corpus were split into 995 segments, which were subjected to further analysis.

All words in the texts were assigned part-of-speech tags employing a tag set CLAWS5, as given in the British National Corpus (Leech & Smith, 2000). Because content words are more likely than function words to capture the ideational content of a text and poetic diction as well as to convey the poets' mindsets, the current study limited its focus to nouns, verbs, and adjectives. Adverbs were excluded from the analysis because some of them do not convey semantic elements; rather, they behave like grammatical elements, which makes it challenging to distinguish them using part-of-speech tags. Once the poems were divided into 320-word consecutive segments, all other words except nouns, verbs, and adjectives were removed in accordance with the part-of-speech tags. The part-of-speech tags for the analysis are displayed in Table 3 with indications from the UCREL CLAWS5 Tagset (https://ucrel.lancs.ac.uk/claws5tags.html) in blackets.

Table 3. Part-of-speech tag list to be analyzed

Nouns	NN0	[Common noun, neutral for number]
	NN1	[Singular common noun]
	NN2	[Plural common noun]
Verbs	VVB	[The finite base form of lexical verbs]
	VVD	[The past tense form of lexical verbs]
	VVG	[The -ing form of lexical verbs]
	VVI	[The past participle form of lexical verbs]
	VVN	[The infinitive form of lexical verbs]
	VVZ	[The -s form of lexical verbs]
Adjectives	AJ0	[Adjective (general or positive)]
	AJC	[Comparative adjective]
	AJS	[Superlative adjective]

The Machine Learning for Language Toolkit was utilized to apply LDA to the segments (McCallum, 2002). The number of topics was determined to be 20, based on the findings from earlier experimental trials, which ranged from 10 to 100. The optimal number of topics was determined by evaluating the LDA results and the close reading of the poems. After reducing the number of candidate topics ranged from 18 to 30, the author repeated the LDA process for each number of topics. Generally, the LDA results vary with each run; however, the 20-topic configuration produced consistent outcomes because the terms assigned to each topic varied minimally. Therefore, the author of this study decided that 20 is the most fitting number of topics for further analysis.

3. Results of the LDA

The LDA output results are discussed below; some results point out topics for further discussion. The observation of the outcome provides the answer to the first research question of this article: 1. Can LDA detect the differences in poetic diction between works by Alfred and Charles? Table 4 presents a subset of the LDA output results, including topic numbers, alpha values, and the most salient keywords associated with each topic. The keywords are arranged in descending order of weight from top left to bottom right. The universality of each topic is indicated by the alpha values: a lower alpha value denotes that the topic appears in fewer segments, while a higher alpha value indicates that the topic appears more prominently across multiple segments in the corpus.

The heatmap representation (Figure 1) provides a visual illustration of the representativeness of the topics, with colors indicating the degree of representation. The darker the cells, the higher the density of the given topic in the poems, and the whiter cells represent a significantly lower topic density. The 20 topics were arranged in a vertical sequence, and the groups of poems (segments) were aligned in a horizontal sequence. The 995 segments for 525 poems have been classified into 12 groups, as it was not feasible to display all 525 works in a single heatmap, given the limitations of a standard sheet of paper. The titles of each group denote the following: A27, Alfred's poems from *Poems by Two Brothers*; ac27, poems with doubtful authors (Alfred or Charles) from *Poems by Two Brothers*; Aup, Alfred's poems included in *Poems by Two Brothers* 1893 edition but not in the 1827 edition; C27, Charles's poems from *Poems by Two Brothers*; T20s, Alfred's poems, which were written in the 1820s but remained unpublished and were not included in the 1827 or 1893 editions of *Poems by Two Brothers*; T30s–T90s, Alfred's poems published or written in the stated range of years. The vertical line titles show the topic numbers with the two most prominent keywords of each topic connected by an underscore (e.g., merry_bride for Topic 0).

As illustrated in the heatmap in Figure 1, the cells representing the most general topics, Topics 5, 12, and 13, are predominantly represented by darker colors. The differentiation between the three topics is based on the frequency of appearance of specific groups. The cells for Topic 5 are observed to be colored darker in the groups of poems published/written by Alfred during the period from the 1840s to the 1890s (T40s–T90s). Topics 12 and 13 exhibit a darker hue in the groups comprising the authors' early works, which were published/written from the 1820s to the 1830s. While Topic 12 demonstrates a darker coloration in the groups where Charles's name is

Table 4. Output result of LDA (Topic number, alpha values, and top 20 keywords)

Topic	Alpha values	Top 20 keywords
0	0.0483	merry_a, bride_n, bridegroom_n, days_n, shake_v, bone_n, cow_n, borne_v, plains_n, praise_n, milking_v, dance_n, whisper_v, green_a, silent_a, grow_v, sits_v, cause_n, goes_v, month_n
1	0.1147	throne_n, war_n, high_a, king_n, woe_n, glorious_a, pride_n, sword_n, proud_a, trumpet_n, fame_n, fiery_a, glory_n, fire_n, course_n, bow_n, battle_n, earthly_a, rise_v, strength_n
2	0.4006	came_v, said_v, heard_v, saw_v, went_v, man_n, knew_v, hand_n, made_v, fell_v, took_v, stood_v, ring_n, dead_a, left_v, spoke_v, turn’d_v, ran_v, isle_n, look’d_v
3	0.0774	death_n, men_n, glory_n, ship_n, battle_n, fight_v, hill_n, sea_n, roof_n, rode_v, arm_v, die_v, right_n, banner_n, fight_n, wives_n, gallant_a, bold_a, devil_n, fought_v
4	0.0613	follow_v, whirl_v, sun_n, winds_n, morning_n, hide_v, forgotten_v, science_n, song_n, slope_n, ends_n, wrong_a, grown_v, woo_v, fine_a, youth_n, earth_n, jewel_n, faint_v, won_v
5	0.7318	heart_n, life_n, love_n, let_v, old_a, know_v, day_n, little_a, love_v, world_n, mother_n, child_n, friend_n, night_n, gone_v, go_v, dead_a, happy_a, good_a, true_a
6	0.0670	pride_n, bones_n, city_n, vale_n, midnight_n, solemn_a, echoing_a, form_n, dark_a, hangs_v, broad_a, cloudy_a, holy_a, grave_n, sombre_a, varied_a, steadfast_a, shades_n, valley_n, branches_n
7	0.0502	time_n, golden_a, good_a, prime_n, old_a, goose_n, honour_n, goodly_a, reason_n, great_a, rhyme_n, worthy_a, side_n, place_n, teach_v, immortal_a, pleasure_n, harder_a, weather_n, stream’d_v
8	0.2223	flowers_n, golden_a, sweet_a, green_a, air_n, year_n, sing_v, song_n, fair_a, summer_n, happy_a, birds_n, flower_n, spring_n, music_n, brook_n, fresh_a, young_a, tree_n, river_n
9	0.0706	glowing_a, god_n, realms_n, fire_n, secret_a, lyre_n, boundless_a, lute_n, lustre_n, maze_n, chords_n, countless_a, bard_n, harp_n, lay_n, roses_n, give_v, sway_n, magic_a, reign_n
10	0.4263	man_n, things_n, power_n, human_a, time_n, nature_n, men_n, life_n, world_n, faith_n, truth_n, mind_n, soul_n, words_n, age_n, great_a, wise_a, make_v, years_n, times_n
11	0.1368	land_n, great_a, queen_n, people_n, men_n, let_v, name_n, freedom_n, king_n, free_a, war_n, voice_n, kings_n, hearts_n, isles_n, cause_n, ancient_a, health_n, friends_n, sea_n
12	0.6745	eyes_n, heart_n, love_n, life_n, death_n, soul_n, sweet_a, tears_n, face_n, light_n, hope_n, fair_a, spirit_n, eye_n, see_v, low_a, mind_n, place_n, full_a, joy_n
13	0.6726	light_n, night_n, earth_n, day_n, sun_n, deep_a, heaven_n, dark_a, bright_a, voice_n, sky_n, sound_n, moon_n, white_a, sea_n, wind_n, stars_n, high_a, eyes_n, cloud_n
14	0.0775	gate_n, lime_n, garden_n, oak_n, seed_n, city_n, boughs_n, eddies_n, maiden_a, folded_a, boat_n, broad_a, windy_a, fern_n, read_v, bridge_n, feed_v, rock_n, beech_n, farm_n
15	0.0677	gods_n, priest_n, bread_n, god_n, fire_n, cross_n, holy_a, flesh_n, saints_n, saved_v, mercy_n, prayer_n, sin_n, mountain_n, saint_n, plague_n, people_n, word_n, hymns_n, leper_n
16	0.0486	let_v, ring_v, form_n, riflemen_n, storm_n, rave_n, cup_n, look_v, rose_n, grave_n, lisette_n, dainty_a, ready_a, green_n, wine_n, fill_v, grave_a, folds_v, order_n, warm_a
17	0.0734	eye_n, ringlet_n, bright_a, charm_n, soft_a, touch_n, sure_a, charms_n, wing_n, view_n, gay_a, gale_n, ecstasy_n, passing_a, scene_n, beauties_n, shrine_n, appear_v, dye_n, virtue_n
18	0.0518	brows_n, seemed_v, smile_n, stood_v, verge_n, smiling_v, flitting_v, gleam_n, bridge_n, clearer_a, frown_n, constant_a, sense_n, floating_a, inward_a, rays_n, wandering_v, idol_n, lamb_n, solid_a
19	0.0476	poet_n, art_n, popular_a, sake_n, muses_n, laurel_n, claim_n, fame_n, gave_v, sow_n, bailiff_n, price_n, college_n, prate_v, fire_n, friends_n, wrong_a, line_n, days_n, grant_v

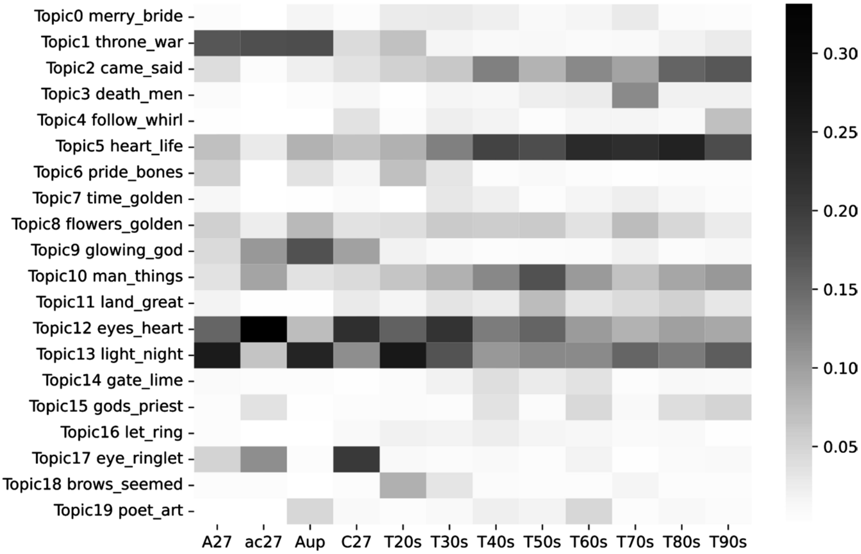


Figure 1. Heatmap of topic densities in clonorogical categories

assigned (namely, ac27 and C27), Topic 13 also comprises dark cells in

Figures 2 and 3 present the top 50 mean density poems for the two topics, which are identified in the preceding paragraph as Topics 17 and 1, respectively. The heatmap in Figure 1 depicts the average densities of the topics for the 12 groups. Figures 2 and 3, in contrast, represent more specific densities for the poems. Given that some poems were split into multiple segments during the LDA and the densities were assigned per segment, the density values for several segments for one poem were calculated and averaged to plot the bar charts.

Topic 17 is a prevalent topic in Charles's poems, as illustrated in Figure 2. A total of 33 poems are assigned to Charles, but 17 Alfred's works, published in various years, are also included in the top 50 poems containing Topic 17. Within 17 Alfred's poems of Topic 17, 11 poems were from the collection of *Poems by Two Brothers* (1827). Of the 11 poems, nine were marked as written by Alfred and two poems were doubtful but assumed to be by Alfred. The six poems discussed above by Ricks are included in the top 50 poems of Topic 1's density (Figure 3). In addition to the six poems, 24 poems were drawn from *Poems by Two Brothers* (1827). Four works were written/published in the 1820s by Alfred, but they are not included in the collection *Poems by Two Brothers*. Excluding the doubtful author works "The Deity" (ac17_ITY), "The Dying Christian" (ac27_IAN), and "Switzerland" (a27_AND), the 41 poems of Topic 1 are

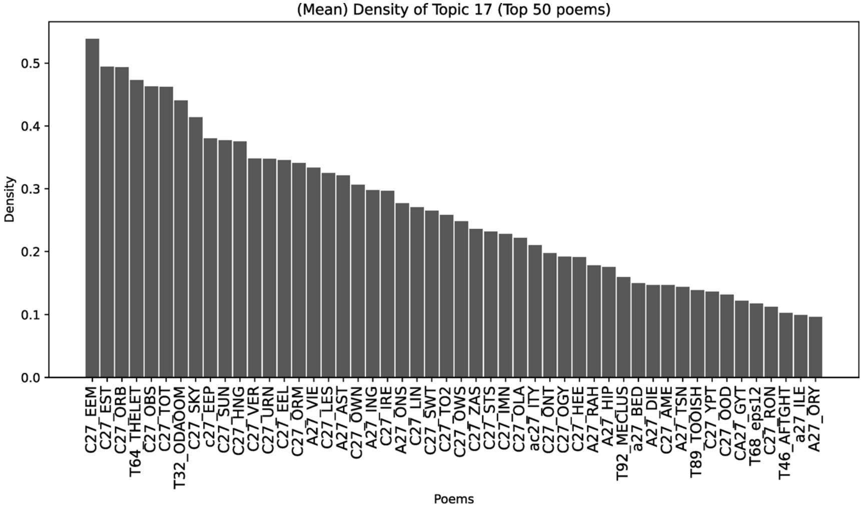


Figure 2. Bar plot of the (mean) density of Topic 17

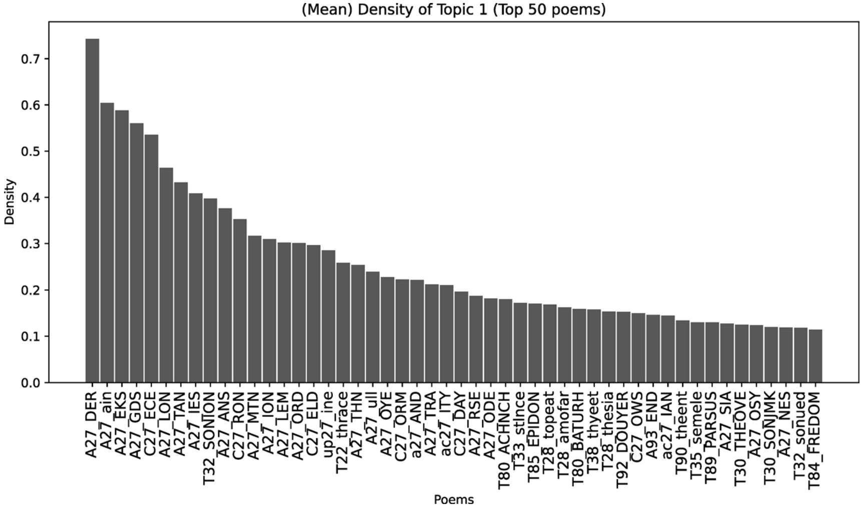


Figure 3. Bar plot of the (mean) density of Topic 1

Alfred's poems.

The results of the LDA analysis indicate that Topic 17 is heavily represented in Charles's poems, whereas Topic 1 is notably present in Alfred's poems from 1827. Furthermore, Topic 9 also emerged in Charles's oeuvre, albeit with a lower frequency than Topic 17. It should be noted that these topics did not exclusively manifest in the works of just one of the brothers. With that acknowledged, the LDA outcomes indicated the tendencies of topics for both Alfred and Charles. In this regard, the initial research question, "Can LDA detect the differences in poetic diction between Alfred's and Charles's works?" can be answered in the affirmative. The following section further narrows the discussion and provides meticulous observation of the two most prevalent topics mentioned in this section. The discussion section of this paper thus attempts to elucidate answers to the second research question, namely, "If LDA detects differences, what characteristics do the two authors exhibit?"

4. Discussion

In this section, the author undertakes an observation and discussion of two topics that were previously identified: Topics 17 and 1. Topic 17 was featured primarily in Charles's poems, whereas Topic 1 is particularly evident in Alfred's poems from 1827. This section is divided into two sections, with each section addressing a specific topic: Topic 17 is discussed in section 4.1 and Topic 1 in section 4.2.

4.1 Topic 17

Topic 17 was a recurring topic in Charles's poems. The terms within the top 20 keywords of Topic 17 are predominantly employed to convey romantic sentiments, feelings toward others (particularly women), and references to women's physical appearance, including the keywords *eye*, *bright*, *charm*, *touch*, *charms*, *ecstasy*, and *beauties*. The noun *ecstasy* is the thirteenth keyword of Topic 17, and is exclusive to the poems of *Poems by Two Brothers*. Alfred used it on two occasions, and Charles utilized it eight times. The poems exude profound romantic passion and ardor, as evidenced by the following lines (boldface added; from this point onward, the use of bold text in quotations will indicate that the referenced word has been assigned to the topic under discussion): "To gaze on thee is **ecstasy**, /Is **ecstasy** — but pain:" ("Oh were This Heart of Hardest Steel," ll. 25–26; C27_EEL) and "Why did I burn with feverish **ecstasy**, /Stung with her scorn, and ravish'd with her praise?" ("The Slighted Lover," ll. 7–8; C27_VER). The fervor of the language and the lines themselves were uniquely present in the 1827 collection, evoking a sense of youthful vigor and association. The poems' content prompted Hallam Tennyson, a son of Alfred Tennyson to say, "As an outburst of youthful poetic enthusiasm, the book is not wanting in interest and a certain charm, although full of the boyish imitation of other poets" (Tennyson H., 1897, p. 22).

Despite the fact that the words assigned to Topic 17 are seen throughout Charles's works, the results revealed that Alfred's poem from 1864, "The Ringlet" (T64_THE-LET), also ranked within the fourth density of Topic 17 (Figure 2). In "The Ringlet," the term *ringlet* appears with notable frequency. The 11th line of the poem states, "My **ringlet**, my **ringlet**," and the 25th line continues with the repetition of "O **Ringlet**, O **Ringlet**." There are 10 instances in the poem where similar lines repeat the term *ringlet* twice in a line bringing the total occurrences of the word to 20. The term *ringlet* is the second keyword within Topic 17, yet it does not appear in Charles's works. A total of 24 instances of the term *ringlet* were identified within the corpus. Of these, 20 occurred within the poem "The Ringlet," while the remaining four were distributed across "The Talking Oak" (1842), *In Memoriam A.H.H.* (1850), and "The Ring" (1889) written by Alfred.

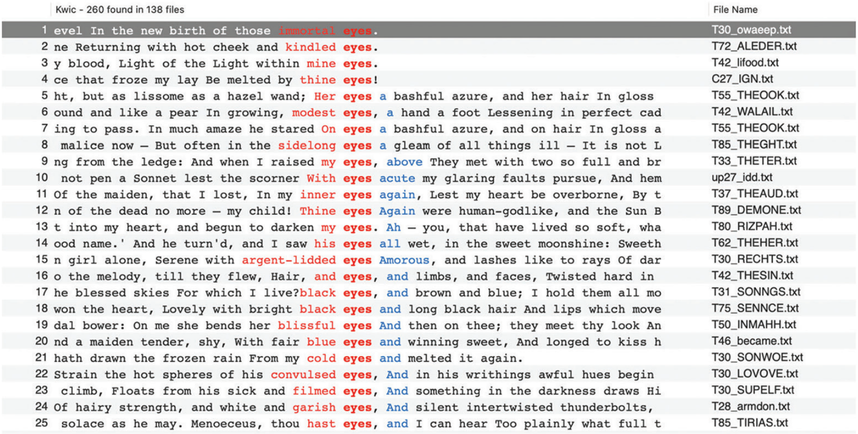
The term *ringlet*'s occurrences were all assigned to Topic 17, although its use is not exclusive to Charles. The third keyword of Topic 17, *bright*, exhibits a distinctive pattern from Topic 17. The total number of occurrences of the word *bright* in the entire corpus is 172, and its frequency it appears in Topic 17 is 19. The 153 instances were allocated to Topics 12 or 13, which are more universal topics than Topic 17, as indicated by their alpha values. Of the 19 appearances of the term *bright* in Topic 17, 13 were found in Charles's works and six in Alfred's. When using LDA, it is not uncommon for the same term to be sorted into different topics. This is due to the fact that LDA utilizes the concept of a "bag of words" concept, whereby the probability of co-occurrence is analyzed within a given segment. If the bags of words exhibit disparate patterns of co-occurrence, the words within the bags can be assigned to distinct topics. The following excerpts illustrate the poems where *bright* appears, where one is assigned to Topic 12 and the other to Topic 17. Alfred's poem, "The Grave of a Suicide" (A27_IDE), is presented on the left, and Charles's poem, "The Slighted Lover" (C27_VER), is presented on the right.

> HARK! how the gale, in mournful notes and stern,
> Sighs thro' yon grove of aged oaks, that wave
> (While down these solitary walks I turn)
> Their mingled branches o'er yon lonely grave!
> ...
> For thou, wed to misery from the womb —
> Scarce one <u>bright</u> scene thy night of darkness knew!
>
> ("The Grave of a Suicide": ll. 1–4, 11–12)

> I LOVED a woman, and too fondly thought
> The vows she made were constant and sincere;
> But now, alas! in agony am taught,
> That she is faithless — I no longer dear!
>
> Why was I frenzied when her **bright** black eye,
> With ray pernicious, flash'd upon my gaze?
>
> ("The Slighted Lover": ll. 1–6)

The term *bright*, which is underlined in Alfred's "The Grave of a Suicide," has been assigned to Topic 12, while the *bright* in Charles's "The Slighted Lover," which appears in bold, has been assigned to Topic 17. Although the excerpts do not display the entirety of each poem's lines, the differences between the two can be discerned. While Alfred's work demonstrates a sense of lamentation pertaining to the life of a person (*yon*), Charles's poem illustrates the sentiment of remorse experienced by the individual (*I*). The term *bright* is employed in two distinct ways in the two poems. In Alfred's poem, it is used to signify both light in the darkness and hope in the context of a person's miserable life. In Charles's poem, however, it is used simply to modify the description of a lady's (*her*) "black eye."

The singular word *eye* was the most significant keyword of Topic 17; it was assigned to Topic 17 in the fifth line of the excerpt of Charles above, where the third keyword *bright* modifies. The frequency of the singular word *eye* was 118 across 80 works, while the frequency of plural *eyes* was 260 across 138 works within the 525 works composing the corpus. The discrepancy between the frequency of the singular and plural forms of the word *eye* does not invalidate the intuitive assumption that the plural form is more prevalent. This is because the human body has two eyes, and the *eye* is defined as "one of the two parts of the body" (s.v. eye, *n*. 1.: *Longman Dictionary of Contemporary English* (2014); underline added by author). Consequently, when the organ is referenced in language describing body parts, the plural form *eyes* is often employed. A concordance line of *eyes* in the corpus is shown in Figure 4 as an example of plural *eyes* usages. The singular form *eye* is not precluded, however, and there may be motivations or reasons for the distinct usages of plural and singular forms of the word in both Charles's and Alfred's works.



```
      Kwic - 260 found in 138 files                                                        File Name
 1 evel In the new birth of those immortal eyes.                                           T30_owaeep.txt
 2 ne Returning with hot cheek and kindled eyes.                                           T72_ALEDER.txt
 3 y blood, Light of the Light within mine eyes.                                           T42_lifood.txt
 4 ce that froze my lay Be melted by thine eyes!                                           C27_IGN.txt
 5 ht, but as lissome as a hazel wand; Her eyes a bashful azure, and her hair In gloss     T55_THEOOK.txt
 6 ound and like a pear In growing, modest eyes, a hand a foot Lessening in perfect cad    T42_WALAIL.txt
 7 ing to pass. In much amaze he stared On eyes a bashful azure, and on hair In gloss a    T55_THEOOK.txt
 8  malice now – But often in the sidelong eyes a gleam of all things ill – It is not L    T85_THEGHT.txt
 9 ng from the ledge: And when I raised my eyes, above They met with two so full and br    T33_THETER.txt
10  not pen a Sonnet lest the scorner With eyes acute my glaring faults pursue, And hem    up27_idd.txt
11 Of the maiden, that I lost, In my inner eyes again, Lest my heart be overborne, By t    T37_THEAUD.txt
12 n of the dead no more – my child! Thine eyes Again were human-godlike, and the Sun B    T89_DEMONE.txt
13 t into my heart, and begun to darken my eyes. Ah – you, that have lived so soft, wha    T80_RIZPAH.txt
14 ood name.' And he turn'd, and I saw his eyes all wet, in the sweet moonshine: Sweeth    T62_THEHER.txt
15 n girl alone, Serene with argent-lidded eyes Amorous, and lashes like to rays Of dar    T30_RECHTS.txt
16 o the melody, till they flew, Hair, and eyes, and limbs, and faces, Twisted hard in     T42_THESIN.txt
17 he blessed skies For which I live?black eyes, and brown and blue; I hold them all mo    T31_SONNGS.txt
18 won the heart, Lovely with bright black eyes and long black hair And lips which move    T75_SENNCE.txt
19 dal bower: On me she bends her blissful eyes And then on thee; they meet thy look An    T50_INMAHH.txt
20 nd a maiden tender, shy, With fair blue eyes and winning sweet, And longed to kiss h    T46_became.txt
21 hath drawn the frozen rain From my cold eyes and melted it again.                       T30_SONWOE.txt
22 Strain the hot spheres of his convulsed eyes, And in his writhings awful hues begin     T30_LOVOVE.txt
23  climb, Floats from his sick and filmed eyes, And something in the darkness draws Hi    T30_SUPELF.txt
24 Of hairy strength, and white and garish eyes, And silent intertwisted thunderbolts,     T28_armdon.txt
25  solace as he may. Menoeceus, thou hast eyes, and I can hear Too plainly what full t    T85_TIRIAS.txt
```

Figure 4. Concordance lines of "eyes" from the corpus

Focusing on the singular word *eye* assigned to Topic 17, the term was observed a total of 26 times across 18 works. Of the 26 instances, 17 were found in Charles's poems, nine were present in Alfred's poems. Among the 17 instances of the *eye* in Charles's works, seven were observed between the lines shown in following (1)–(7), while the remaining ten instances were found at the end of the lines. In excerpts (5) and (7), *eye*-s are referred to as a singular entity due to the grammatical requirements of the language. In line (5), the preceding adjective *each* demands a singular noun form. In

(7), the noun *eye* is used adjectivally to modify the noun *beam*; and it is singular because nouns are usually singular in adjectival use. The *eye* in lines (5) and (7) is therefore an irreplaceable unit, whereas the other *eye* can be replaced with a plural form. In the instances of the other five uses of *eye*-s, a metonymic usage can be observed, whereby the singular *eye* denotes the entire body part, namely two eyes. In excerpts by Alfred, seven out of the nine instances of *eye* assigned to Topic 17 occurred in the middle of the lines. All seven instances represent the metonymic *eye*.

(1) But he whose **eye** the light can chase,
(“Borne on Light Wings of Buoyant Down”: l. 17)
(2) The **eye** with wonder gazes there, (“The Stars of Yon Blue Placid Sky”: l. 5)
(3) Mocks the foil’d **eye** that would its hues arrest,
(“The Dew with which the Early Mead is Drest”: l. 3)
(4) That **eye**, that cheek, that lip, possess
(“Oh were this Heart of Hardest Steel”: l. 5)
(5) The lightning too each **eye** in dimness shrouds, (“The Thunder-storm”: l. 13)
(6) The **eye** must catch the point that shows, (“Lines”: l. 1)
(7) I may not see the glazed **eye** beam;
(“Still Mute and Motionless She Lies”: l. 30)

Considering the position in which the term *eye* occurs in the Topic 17 poems, it is notable that the most frequent instances are at the ends of lines. A total of 26 instances of the term *eye* have been identified within Topic 17. Of these instances, 10 occur at the end of lines and constitute part of the foot rhymes in the poems of Charles and two in Alfred, respectively. The two cases of the line-end *eye* in Alfred’s poems both rhyme with the word *sky*. In Charles’s oeuvre, the rhyming partner terms are diverse, including *sky*, *die*, *fly*, and *dye*. The boldface terms in following excerpts illustrate the instances in Charles’s poems where the terms *eye* and *dye* rhyme. The term *dye* was identified as the 19th most significant keyword of Topic 17. The word *dye* only appeared exclusively in Charles’s works.

O’er her sweet cheek’s once lovely **dye**,
  I shudder’d as I turn’d

From the sad spot, and in mine **eye**
  The full warm tear-drop burn'd.
(“A Sister Sweet Endearing Name” (C27_AME): ll. 17–20; bold added)

But winter came — its varied **dye**
Each morn grew fainter to mine **eye**;
Till, with'ring, it was bright no more,
Nor bloom'd as it was wont before:
(“Still Mute and Motionless She Lies” (C27_STS): ll. 13–16; bold added)

Given that the pronunciation of *eye* is comprises of a single diphthong, /aɪ/, it can be inferred that the entirety of the word *eye* itself, or the entire sound of the word /aɪ/, represents the target for rhyming with another word. Departing from the Topic 17 elements and contemplating Charles's oeuvre in a more comprehensive manner, however, an intriguing suggestion emerges. The act of rhyming is typically understood to entail the utilization of identical or analogous vocal elements at the end, beginning, and/or middle of poetic lines. The same or similar sounds are based on vowels, and it is not necessary for the consonants preceding or following the vowel to be identical. Additionally, as the term *similar* indicates, the vowel (and consonant) sound(s) need not be an exact match. With this established, in Charles's poems, the exact match of a vowel and the subsequent consonant(s) frequently occurs: for example, in *pow'r*/*flow'r*, *roll*/*pole*, *fire*/*ire* (“In Summer when All Nature Glows” (C27_OWS): ll. 19–24); in *stage*/*age*, *view*/*woo*, *awake*/*take*, *steals*/*heels*/*reveals* (“Still Mute and Motionless She Lies”: ll. 1–9). As previously stated, Charles's “Still Mute and Motionless She Lies” provides an illustrative example of *eye/dye* rhyming. It commences with a rhyme involving the plural *eyes*:

STILL, mute, and motionless she lies,
The mist of death has veil'd her eyes.
And is that bright-red lip so pale,
Whose hue was freshen'd by a gale
More sweet than summer e'er could bring
To fan her flowers with balmy wing!
(“Still Mute and Motionless She Lies”: ll. 1–6; underline added)

A comparison of the *eyes* in the second line and *eye* in the 14th line (as seen in the previous excerpt) of the "Still Mute and Motionless She Lies" reveals a distinct contrast between the two lines. The contrast hinges on the pronoun used to describe the *eye*/*eyes*: the genitive case third-person pronoun *her* or the first-person possessive pronoun *mine*. Upon expanding our "eye" to include the anteroposterior lines, terms rhyming with *eye*/*eyes* are observed to differ between the lines. Given that the rhymes in Charles's work frequently align with both vowel and subsequent consonant sounds, it is reasonable to hypothesize that the use of singular or plural nouns in analogous positions may be influenced by rhyme or sound structures.

Similar to *eye*/*eyes*, the singular and plural forms of *charm* were among the top 20 keywords of Topic 17. The singular form of *charm* was found three, nine, and eight times in Alfred's works from 1827, in Alfred's works from the 1830s to the 1890s, and in Charles's works, respectively. The plural form of *charms* appeared two, three, and nine times in Alfred's works from 1827, in Alfred's works from the 1830s to the 1890s, and in Charles's works. In Charles's oeuvre, there is only one instance of *charms* occurring at the end of a line, where it rhymes with *warms* ("Imagination" (C27_IMN): ll.15–16), and otherwise, the word is found in the middle of the lines. Conversely, the three instances of Alfred's poems are positioned at the end of the lines, and all three uses of the word *charms* rhyme with *arms*. The aforementioned examples of the distinctions between the singular and plural forms of *charm* indicate that there is not a single, straightforward reason or motivation underlying the differences in usage. Nevertheless, the observations on singular and plural differences in one topic indicated the possibility that sound preferences might be a contributing factor in the alteration of their forms. It is regrettable that LDA is unsuitable for the analysis of sound and grammatical elements. It is therefore not possible to conclude that LDA has identified the rhyming preferences of the author(s). To ascertain the rhyme and/or sound structures and preferences of the both Charles and Alfred, further analyses employing optimal methods are required in future studies.

4.2 Topic 1

Topic 1 appeared significantly in the *Poems by Two Brothers*. Of the top 50 density poems in Figure 3, 30 were included in the 1827 collection. In addition to the 30 poems from 1827, five other poems were written by Alfred in the 1820s. Additionally, the 43 poems among the top 50 poems of Topic 1 are also Alfred's poems. Topic 1 can therefore be considered a topic that primarily represents Alfred's poems, particularly those written during his early career. Unlike the poems of Topic 17, the poems in which Topic 1 frequently appears tend to primarily address masculinity or substances, evoking images of men and scenes in which men are often depicted. Of the top 20 keywords of Topic 1, the fourth keyword, *king*, was directly related to the concept of a male crown. Other keywords, including *throne*, *fame*, and *bow*, are associated with notions of nationhood and royal authority. The keywords *war*, *sword*, *fire*, *battle*, and *strength* are linked to both nations and masculinity, as historically, men have been the ones to serve their nations or crowns. Wars or battles are often initiated for the purpose of protecting or expanding a nation, region, or diadem. Further observation revealed additional relationships between keywords. The terms *glorious*, *pride*, *proud*, *fame*, *glory*, and *trumpet* are strongly associated with the keywords *war* and *battle*. These associations are evident in various poems by Alfred that appear in the 1827 collection. The following quotations are "The High Priest to Alexander" (A27_DER) and "Exhortation to the Greeks" (A27_EKS), the first and third density poems of Topic 1.

> **Go forth**, thou man of force!
>   The world is all thine own;
> Before thy **dreadful course**
>   Shall **totter** every **throne**.
> Let India's **jewels** glow
>   Upon thy **diadem**:
> Go, forth to **conquest go**,
>   But **spare** Jerusalem.
>   For the **God** of **gods**, which **liveth**
>   Through all **eternity**,
>   'Tis He alone which **giveth**
>   And **taketh victory**:
>
> ("The High Priest to Alexander": ll. 1–12)
>
> **AROUSE** thee, O Greece! and **remember** the day,
> When the millions of Xerxes were **quell'd** on their
>   way!
> **Arouse** thee, O Greece! let the **pride** of thy name

Awake in thy bosom the light of thy **fame**!
. . .
Remember each day, when, in **battle array**,
   From the fountain of **glory** how largely ye **drunk**!
For there is not aught that a freeman can **fear**,
   As the **fetters** of **insult**, the name of a **slave**;
And there is not a voice to a nation so dear,
   As the **war-song** of **freedom** that calls on the **brave**.
(“Exhortation to the Greeks”: ll. 1–4, 21–26)

As evidenced by the aforementioned poems, the top 20 keywords of Topic 1 can be discerned not only in isolation but also in conjunction with their synonyms and related terms assigned to the topic. In “The High Priest to Alexander,” the term *diadem* is associated with the crown of a nation. Additionally in this same poem, the term *victory* is related to the concepts of *glory*, *war*, and *battle*, which are among the top 20 keywords of Topic 1. In “Exhortation to the Greeks,” *quell’d*, *slave*, *war-song*, and *freedom* are correlated with the concept of *battle*. It is apparent that the locations of these battles and wars were not necessarily within the boundaries of the United Kingdom, as illustrated by references to *Jerusalem* and *Greece* in “The High Priest to Alexander” and “Exhortation to the Greeks.” Furthermore, the top 20 keywords of Topic 1, as well as the poems themselves, demonstrate a sense of masculinity or vigor, despite the paucity of references to individuals in the poems.

The following excerpt is from “Written During the Convulsions in Spain” (A27_ain), the second highest density poem of Topic 1, written by Alfred. In this poem, the top 20 keywords, as well as the words related to the keywords, such as *arm*, *combat*, and *fight*, were observed as the terms of Topic 1. In addition, the term *heroes*, connoting masculinity, was identified.

Strong be their **arm** in **war**,
Brilliant their **glory**’s star,
  Fierce be their **valour** and **fearful** their name!
. . .
Where are thine **heroes** hid?

Arm them for **combat** and shout, ‘To the **fight**!’
Shake the **throne** of thy Lord
To its **base** with their **sword**,
So, on to the **combat**, and God **help** the right!
(“Written During the Convulsions in Spain”: ll. 16–18, 32–36; bold added)

Masculinity also emerges in Charles’s poems, yet these evince disparate emotional nuances compared to those expressed by Alfred. The following quotes from Charles’s poems are not as vigorous as Alfred’s ones, whereas several Topic 1 keywords are employed in Charles’s poems. In the excerpt of “On the Death of Lord Byron” (C27_RON), the terms *hero*, *career*, *blaze*, and *fame*, are designated as the Topic 1 keywords. In this poem, the singular *hero* refers to George Gordon Byron (1788–1824), who is regarded as a representative poet of the Romantic era. Despite the absence of any explicit references to warfare or combat, the poem’s principal subject is a male figure. While previous literature, such as by Shaw (1973) and Thomas (2019), has indicated similarities in the poetic styles of Alfred and other Romantic poets (for example Percy Bysshe Shelley, John Keats, and William Wordsworth), there is a paucity of lamentations in Alfred’s poetry regarding the loss of other Romantic poets and a dearth of commentary on the works of other Romantic poets. In “On the Death of Lord Byron,” however, Charles expresses great fervor in his lamentation of the loss of Lord Byron. Although the existence of “On the Death of Lord Byron” does not directly refute the notion that Alfred held other poets in high regard, it does imply that Alfred’s respect for and engagement with other poets may not be as profound and deep as Charles’s. The enthusiasm respectively displayed by Alfred and Charles diverged during their adolescence, however. Charles’s poem “The Battle-field” (C27_ELD) depicts a scene of battle or war but does not include the terms *battle* or *war*. Despite this, terms related to warfare or battle are ascribed to Topic 1, including *chaos*, *contest*, *madden*, *trumpet*, *barbarous*, *bray*, and *cannon*. Likewise, the use of the term *heroes* in “The Battle-field” suggests that the poem also extols masculinity.

THE **hero** and the bard is gone!
His bright **career** on earth is done,
Where with a comet’s **blaze** he shone.

. . .
Was Byron's hope — was Byron's aim:
With ready heart and hand he came;
But perish'd in that path of **fame**!

("On the Death of Lord Byron": ll. 1–3, 37–39)

THE heat and the **chaos** of **contest** are o'er,
To mingle no longer — to **madden** no more:
And the cold forms of **heroes** are **stretch'd** on the
plain;
Those lips cannot breathe thro' the **trumpet** again!
. . .
I — heard, oh! I heard, when, with **barbarous bray**,
They leapt from the mouth of the **cannon** away;

("The Battle-field": ll. 1–4, 9–10)

While a handful of Charles's poems are included in the top 50 poems of Topic 1, the number of Alfred's poems in the topic was significantly greater. It can be concluded that LDA identified the predominantly male elements, masculinity, and enthusiasm for and in aspects of battle in both Charles's and Alfred's poems in Topic 1, appearing more frequently however in works by Alfred.

5. Conclusion

This study employed the quantitative approach LDA to identify the characteristic diction of Alfred Tennyson and his brother, Charles, in their first publication, *Poems by Two Brothers* (1827). The LDA outcomes indicated that two topics, Topics 17 and 1, were particularly prevalent in the collection. Furthermore, Topic 17 was identified as a more prominent feature in the poems of Charles, while Topic 1 was observed to be a significant element in Alfred's poems. Topic 17 was found to represent terms associated with romantic sentiments directed toward women and descriptions of their physical appearances. In contrast, Topic 1 represented lexical items associated with masculinity, enthusiasm, and battles. The distinction between the two topics suggests that there are differences in the vocabulary and pattern of expression used by Alfred and Charles. Topics 17 and 1 yielded responses to our initial research question, "Can LDA detect the differences in poetic diction between Alfred's and Charles' works?" Upon examination of these topics, we were also able to ascertain answers to our secondary research question, "If LDA detects differences, what characteristics do the two authors exhibit?"

Previous authorship attribution studies have employed neither the LDA nor the analysis of content words, which were the focus of this study. The objective of the present study was not to ascertain the efficacy of LDA in authorship attribution but to identify internal evidence of the distinguishing characteristics of the two brothers. Indeed, the results indicated that the differences between the authors could not be fully classified. The findings of this study indicate the limitations of LDA in terms of achieving complete certainty regarding authorship attribution. However, the accuracy of author estimation can be further enhanced by combining the results of other function words and examining content words using LDA in the context of lyrical poetry studies. This is because quantitative lyrical poetry studies are confronted with the challenge of handling shorter and smaller data than prose text and gaining reasonable data size and results.

Section 4.1, above, posits the potential influence of sound preferences based on the observation of the themes of Topic 17. It would be beneficial for future studies to consider the stylistic features of the poems, including function words and rhymes, to gain insight into the authorship of the poems. Nevertheless, the distinctive diction indicated in this article will constitute an element of internal evidence. Concerning diction, future studies are needed to observe a greater number of poems and topics in order to identify other possible features of each poet. The integration of internal and external evidence, in addition to qualitative and quantitative approaches, will surely further advance the authorship attribution and the study of Alfred's and Charles's poems.

Note

[1] This study employs the reprinted edition of *Poems by Two Brothers* (1827), printed in London by W. Simpkin and R. Marshall, and J. and J. Jackson. The reprinted edition was published by Thomas Y. Crowell in New York, though the precise date of publication is not indicated.

Acknowledgment

This study is supported by a grant from the Japan Society for the Promotion of Science (JSPS) KAKENHI program (grant number 24K22483). Expressions of gratitude are extended to the anonymous reviewers who have provided constructive feedback, leading to significant enhancements in the quality of this article. It should be noted that the responsibility for any residual errors remains with the author.

Bibliography

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Brimley, J. R. (1972). *Tennyson & His Poetry.* New York: AMS Press.

Fujita, I. (2020). "He triumphs; maybe, we shall stand alone": Using correspondence analysis to investigate modal adverbs in Tennyson's poems. *Japanese Association of Digital Humanities 2020*, 21–26.

Fujita, I. (2022). On segment size in poetry analysis using the latent dirichlet allocation method. *Digital Humanities, 3*, 3–15.

Fujita, I. (2023). 'To spread the Word by which himself had thriven': Analysis of Alfred Tennyson's use of language based on the LDA topic model. *English Corpus Studies, 30*, 3–26.

Jockers, M. L. (2014). *Text Analysis with R for Students of Literature.* Heidelberg, New York, Dordrecht, London: Springer Cham.

Leech, G., & Smith, N. (2000, March 17). *Manual to accompany: The British National Corpus (Version 2) with Improved Word-class Tagging*. Retrieved November 25th, 2024 from https://ucrel.lancs.ac.uk/bnc2/bnc2postag_manual.htm

Longman Dictionary of Contemporary English. (2014). Retrieved November 25th from Longman Dictionary of Contemporary English: https://www.ldoceonline.com/

McCallum, A. K. (2002). *MALLET: A machine learning for language toolkit*. Retrieved November 25th from http://mallet.cs.umass.edu

Paden, W. D. (1964). The Tennysons' *poems by two brothers* (1827) reconsidered. *The Library*, *5*, 147–161.

Ricks, C. (1987). *The Poems of Tennyson* (2nd ed.) (Vol. III). London: Longman.

Shaw, D. W. (1973). *Tennyson's Style.* Ithaca: Cornell University Press.

Tabata, T. (2018). Mapping Dickens's novels in a network of words, topics, and texts. *Text mining and Digital Humanities*, *2017*, 47–58.

Tennyson, A.; Tennyson, C. (n.d.). *Poems by Two Brothers*. New York: Thomas Y. Crowell.

Tennyson, A.; Tennyson, C. (1893). *Poems by Two Brothers* (2nd ed.). London: Macmillan.

Tennyson, A. (2013). *Delphi Poets Series Alfred, Lord Tennyson.* East Sussex: Delphi.

Tennyson, H. (1897). *Alfred Lord Tennyson: A Memoir by his Son.* London: Macmillan.

Thomas, J. (2019). *Tennyson Echoing Wordsworth.* Edinburgh: Edinburgh University Press.

UCREL CLAWS5 Tagset. (n.d.). Retrieved November 25th, 2024 from CLAWS part-of-speech tagger for English: https://ucrel.lancs.ac.uk/claws5tags.html

（Iku Fujita, currently at Kyushu University; formerly at Hijiyama University,
E-mail: tnnysn.annie.if@gmail.com）

「研究ノート」

日英・英日パラレルコーパス検索ツール『パラレルリンク』（Ver. 2.0）—インターフェース，検索機能，活用研究などについて—

仁科　恭徳・赤瀬川史朗

Abstract

In this paper, we introduce the concrete interface, various on-board corpora, search functions, implemented statistical processing, and expected applications of Parallel Link (Ver. 2.0), an analysis tool for Japanese-English and English-Japanese parallel corpora that has finally been completed. In particular, four new parallel corpora, BSD, Coursera, Hiragana Times, and JENAAD, have been added to the database, bringing the total number of parallel corpora to 13 and increasing its volume by around 1.3 times. An overview of these added corpora, as well as those to be added in the future, will be presented in detail. The new functions added in Ver. 2.0, such as the context display function, the distribution graph function for each corpus of collocations, the voice function, and the Google Translate link function, will also be introduced.

１．はじめに

仁科・赤瀬川（2021, 2022a, 2022b），仁科（2023）では，これまでに構築された日英・英日パラレルコーパスやその検索ツール，そしてこれらを活用した研究を網羅的に振り返り，2020年度から2022年度にかけて筆者らが開発した日英・英日パラレルコーパスオンライン検索ツール『パラレルリンク』（Ver. 1.0〜1.2）に搭載した9種のパラレルコーパスの概要と，それらを再整備する上で施したテキスト処理やアノテーション，全文検索インデックスの作成，ファイル整理などの一連の作業について詳説した。本稿では，2024年7月下旬に公開された当該ツールの最新版Ver. 2.0に搭載したコーパスや具体的なインターフェース，検索機能，実装されている統計処理，想定される活用研究などにつ

いて紹介する[1]。

2.『パラレルリンク』（Ver. 2.0）のインターフェースと検索機能

2.1. 概要

Munday（2016）は, 翻訳学研究の中でも翻訳補助（translation aids）を促すツールの開発や利用は応用翻訳学（Applied Translation Studies）に含まれる分野であるとし，この中には機械翻訳を含むソフトウェア開発やクラウド・ソーシングを含む共同作業（collaboration），辞書やオンライン検索，パラレルコーパスの参照などが含まれるとする。『パラレルリンク』のプロトタイプの開発経緯については既に仁科・赤瀬川（2021, 2022）で述べたとおりであり，元々は翻訳作業や辞書編纂の手助けとなるような参照ツールを目指して開発した。Ver. 1.0〜1.2 では，現在までに無償公開された 9 種の日英・英日パラレルコーパスを搭載し，オンライン上で網羅的に串刺し検索できることを可能にした。各ジャンルの語数にばらつきはあるものの，今後 KWIC 検索などの検索機能を充実させることで，簡易的な一般参照パラレルコーパスとして活用することも期待できる。また，起点言語（日本語）の検索語に関する精緻な語彙プロファイルの獲得や，その目標言語（英語）における信頼性ある翻訳例が参照できる有益なツールとなりうる。

そこで，本節では，当該ツールの最新版となる Ver. 2.0 に実装されているパラレルコーパスとインターフェースおよび検索機能を紹介する。特に，見出し検索，パタン表示機能，コロケーション表示機能，翻訳表示機能，文脈表示機能，コロケーション分布グラフ，Google 翻訳リンク機能などを紹介し，その後，想定される活用研究や Ver. 3.0 以降の開発計画など，今後の展望についても具体的に示す。

本ツールは，国立国語研究所と Lago NLP（旧 Lago 言語研究所）が開発したブラウザベースのコーパス検索ツール Lago Word Profiler（以下，LWP）（パルデシ･赤瀬川，2011）が原型となっており，「見出し語単位で検索，コロケーションなどを文法項目に分類して整理して表示」することができ（染谷・赤瀬川・山岡，2011），日本語原文の英訳もパタンやコロケーション別に表示させることが可能である。また，ジャンル別の頻度表示や絞り込み機能も併せ持つ。

2.2. レキシカルプロファイリングのツール開発について

仁科・赤瀬川（2021, 2022a）では，テクスト処理，アノテーション，テクストクリーニング，エンコーディングの統一，全文検索インデックスの作成，ファイル整理などについて紹介した。インターフェースの開発には他にもレキシカルプロファイリングデータベースの構築が必要となる。

参考までに，日本語の内容語（動詞，名詞，形容詞，副詞）のレキシカルプロファイリングデータベースを構築する作成手順を簡単に紹介する。まず，フォーマットを統一したコーパス群を CQL で検索できる検索システムを構築する。次に，各内容語の文法パタンを係り受け関係のあるものから抽出できるように CQL クエリーで表現する。抽出した文法パタンは，RD(B)MS（Relational DataBase Management System）の MySQL データベースに格納している。インターフェース開発までの一連の経緯をまとめると，係り受け解析，内容語のパタン作成（名詞，動詞，形容詞（イ形容詞，ナ形容詞），副詞，連体詞），レキシカルプロファイリングデータベースの構築（内容語ごとの共起関係のデータ（共起語，頻度，MI スコア，LogDice, T スコア，対数尤度比））となる。

なお，レキシカルプロファイリングデータベースは，見出し語情報テーブル，文法パタンテーブル，コロケーションテーブル，用例テーブルから構成される。見出し語情報テーブルは，見出し語（漢字かな交じり，ひらがな，ローマ字），品詞（大分類，中分類），頻度など，見出し語に関する基本情報を収録したテーブルである。文法パタンテーブルは，見出し語ごとに，文法パタン別の頻度，パタングループにおける割合などを収録したテーブルである。コロケーションテーブルは，見出し語と文法パタンをキーとして，該当するコロケーションと統計値（前述の単純頻度，MI スコア，LogDice, T スコア，対数尤度比），コーパスごとの頻度を収録したテーブルである。用例テーブルは，見出し語と文法パタンとコロケーションをキーとして，該当する用例と対訳，出典元のコーパス名を収録したテーブルである。

2.3. Ver. 2.0 における搭載コーパスの追加・変更

Ver. 2.0 から新たに追加したパラレルコーパスは，ビジネスシーン対話対訳コーパス The Business Scene Dialogue Corpus（以下 BSD）（https://github.com/tsuruoka-lab/BSD），オープンコースウェア対訳コーパス Coursera Parallel Corpus（以下 Coursera）（https://nlp.ist.i.kyoto-u.ac.jp/EN/?Coursera+Parallel+Corpus），ひらがなタイムズ日英対訳コーパス（以下 Hiragana Times），日英新聞記事対応

付けデータ The Japanese-English News Articles Database（以下 JENNAD）の 4 種である。

BSD は，会議，交渉，雑談といったビジネスシーンに沿った日本語および英語での会話のシナリオを作成し，英語および日本語へ翻訳したコーパスである（中澤・李・Rikters, 2021）。Coursera は教育分野のコーパスであり，世界中の大学のコースを無償でオンライン提供している Coursera（https://www.coursera.org/）から抽出した日英対訳データから成る。これら 2 種のパラレルコーパスについては無償であり，前者はビジネス関連で話し言葉の性質を持ったコーパスであること，後者は教育関連コーパスであることから Ver. 1.0 に搭載した 9 種のパラレルコーパスには不足・欠落している分野・モードであるため追加した。

また，Hiragana Times は今回新たに購入した有償コーパスで，ひらがなタイムズ誌面の対訳データから構成される。1988 年～2017 年までの計 349 冊/ファイル（対訳数 212,230 文）と別冊書籍 Hiratai-Books シリーズ 19 冊/ファイル（対訳数 21,796 文）で構成されており，アカデミック割引でも 100 万円程度必要となる有償ライセンス形式のコーパスである。そして JENAAD は読売新聞と Daily Yomiuri の記事を収録したパラレルコーパスで，公開当時は最大規模のパラレルコーパスであった。長年の間，言語研究者に愛用されてきたパラレルコーパスの雄である（仁科，2014）。これら 2 種のパラレルコーパスについて，前者は有償で購入するために研究費を獲得する必要があったこと，後者は版権の確認など時間を要したことから Ver. 1.0 には搭載できなかったが，今回 Ver. 2.0 でようやく搭載することができた。Ver. 1.2 から Ver. 2.0 にかけて，語数換算で日本語が約 1.35 倍，英語が 1.64 倍，対訳対数が 1.31 倍に増強している。

JESC，Open Source，Reuters，Taiyaku に用いられた類似度フィルタリングには，Google の大規模言語モデル universal-sentence-encoder-multilingual を利用し，異言語（日本語－英語）のセンテンス間の類似度を判定している。また，Ver. 3.0 以降で JParaCrawl から 100 万対程度，Wikimedia から 14 万対程度を追加し，全対訳数 500 万対を目指す。JParaCrawl については，類似度フィルタリングと有害コンテンツの排除を行なった上でデータベースに追加予定である。このコーパスはネットから対訳を探したデータであることから，有害コンテンツなど質の悪いデータが一定数混入しているが，この問題を解決し質の良い用例だけを足すことができれば本ツールで検索できる全データサイズも現在の数倍となり，質・量共にパワーアップする。データ量が増えることで，翻訳ユニットの抽出精度も格段に向上することが期待される。Wikimedia Parallel Corpus に

表 1. Ver 2.0 搭載コーパスに関するまとめ（新規追加，変更点など）

パラレルコーパス	翻訳方向	ステータス	語数（日本語）	語数（英語）	対訳対数
ビジネスシーン 対話対訳コーパス（BSD）	英→日 日→英	新規追加	314,785	277,267	24,578
オープンコースウェア 対訳コーパス（Coursera）	日→英	新規追加	1,147,393	1,036,545	63,118
ひらがなタイムズ日英対訳 コーパス（Hiragana Times）	日→英	新規追加	4,725,779	4,279,484	261,770
日英新聞記事対応付けデータ （JENNAD）	日→英	新規追加	5,287,223	4,957,273	192,763
日英サブタイトルコーパス （JESC）	英→日 (一部，日→英)	類似度フィルタ リング	7,479,726	6,785,436	1,061,623
日英法令対訳コーパス （LAW）	日→英	各用例に法律名 を付与	12,991,680	13,092,136	424,826
大規模オープンソース 日英対訳コーパス （OPENSOURCE）	英→日	類似度フィルタ リング	4,492,098	3,371,492	246,137
ロイター日英記事の対応付け （REUTERS）	英→日	類似度フィルタ リング	2,066,014	1,714,672	76,295
SCoRE 用例コーパス （SCoRE）	英→日	トピック別用例 の追加	213,195	136,302	15,696
日英対訳文対応付けデータ （TAIYAKU）	英→日 (一部，日→英)	類似度フィルタ リング	1,914,472	1,431,563	118,084
Tatoeba 日英対訳コーパス （TATOEBA）	日→英	最新版より用例 追加	2,604,111	2,097,640	279,985
TED Talk 英日コーパス （TED）	英→日	パラレルコーパ スの質を改善	8,937,299	7,654,440	588,956
Wikipedia 日英京都関連文書対 訳コーパス（WIKIPEDIA）	日→英	変更なし	9,436,567	9,940,272	465,955
		Total 2.00 版	52,673,043	56,774,522	3,230,830
		1.20 版	38,934,507	34,646,840	2,459,913

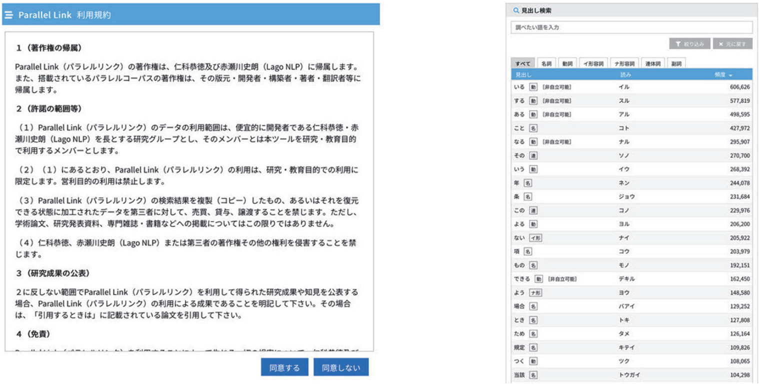
ついては，https://opus.nlpl.eu/wikimedia/ja&en/v20230407/wikimedia を参照されたい。

2.4. Ver. 2.0 のインターフェースと検索機能

本節では，本ツール（https://www.parallellink.org/）の具体的なユーザーインターフェースについて紹介する。本ツールは LWP を基盤として開発されたことから，他のコーパスを搭載した LWP である NINJAL-LWP for BCCWJ（NLB）（プラシャント・赤瀬川，2011）や LWP for ParaNews（現在は公開終了），

Wikipedia-Kyoto LWP（現在は公開終了）を使用したことのあるユーザーにとっては，馴染みのあるインターフェースである。但し，今回は白を基調とした明るく眼に優しいデザインに変更しフォントも見やすくなったことから，旧インターフェースより視認性が高くなっている。なお，上記ウェブサイトからユーザーガイドをダウンロードすることができるので，基本的な検索機能についてはそちらを参照されたい。また，インターフェースやユーザーガイドについては英語版も用意している。画面右上に配置された Display Language をクリックすれば，表示言語を Japanese（日本語）と English（英語）から選択できる。

まず，Home 画面には本ツールに関する簡単な紹介，ユーザーガイド，引用方法，開発チームメンバー，更新履歴，謝辞，問い合わせ先が記載されている。今後，『パラレルリンク』を活用した研究事例の情報なども随時こちらにアップしていく予定である。その Home 画面右上の search をクリックすると，図 1 左の利用許諾のポップアップが表示される。「同意する」を選択すると，図 1 右に示す検索画面に移動する。Ver. 2.0 の段階では日→英方向のみの検索が可能であるため，日本語の各見出し語の情報のみが一覧表示される（Ver. 3.0 以降で英→日検索も可能とする予定）。予め用意された品詞タブは，「すべて」，「名詞」，「動詞」，NLB とは異なる「イ形容詞」と「ナ形容詞」，「連体詞」，「副詞」の計 7 種である。Ver. 2.0 から，「すべて」に掲載されている見出し語には品詞ラベル（例．動）と，補助動詞としても使われる動詞には［非自立可能］のラベルが付与されている。検索したい特定の品詞タブをクリックすれば，各品詞に該当する語の情報が頻度順で表示される。検索ボックスへの文字入力は，か



1（著作権の帰属）

Parallel Link（パラレルリンク）の著作権は、仁科恭徳及び赤瀬川史朗（Lago NLP）に帰属します。また、搭載されているパラレルコーパスの著作権は、その版元・開発者・構築者・著者・翻訳者等に帰属します。

2（許諾の範囲等）

（1）Parallel Link（パラレルリンク）のデータの利用範囲は、便宜的に開発者である仁科恭徳・赤瀬川史朗（Lago NLP）を長とする研究グループとし、そのメンバーとは本ツールを研究・教育目的で利用するメンバーとします。

（2）（1）にあるとおり、Parallel Link（パラレルリンク）の利用は、研究・教育目的での利用に限定します。営利目的の利用は禁止します。

（3）Parallel Link（パラレルリンク）の検索結果を複製（コピー）したもの、あるいはそれを復元できる状態に加工されたデータを第三者に対して、売買、貸与、譲渡することを禁じます。ただし、学術論文、研究発表資料、専門雑誌・書籍などへの掲載についてはこの限りではありません。

（4）仁科恭徳、赤瀬川史朗（Lago NLP）または第三者の著作権その他の権利を侵害することを禁じます。

3（研究成果の公表）

2に反しない範囲でParallel Link（パラレルリンク）を利用して得られた研究成果や知見を公表する場合、Parallel Link（パラレルリンク）の利用による成果であることを明記して下さい。その場合は、「引用するときは」に記載されている論文を引用して下さい。

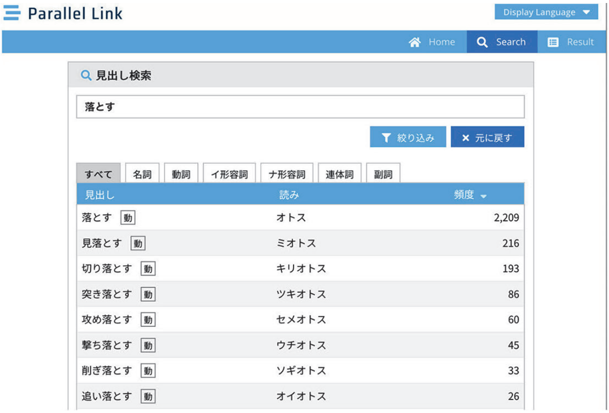
4（免責）

見出し	読み	頻度 ▾
いる 動 ［非自立可能］	イル	606,626
する 動 ［非自立可能］	スル	577,819
ある 動 ［非自立可能］	アル	498,595
こと 名	コト	427,972
なる 動 ［非自立可能］	ナル	295,907
その 連	ソノ	270,700
いう 動	イウ	268,392
年 名	ネン	244,078
条 名	ジョウ	231,684
この 連	コノ	229,976
よる 動	ヨル	206,200
ない イ形	ナイ	205,922
項 名	コウ	203,979
もの 名	モノ	192,151
できる 動 ［非自立可能］	デキル	162,450
よう ナ形	ヨウ	148,580
場合 名	バアイ	129,252
とき 名	トキ	127,808
ため 名	タメ	126,164
規定 名	キテイ	109,826
つく 動	ツク	108,065
当該 名	トウガイ	104,298

図 1. 利用規約のページ（左）と見出し検索の画面（右）

な漢字交じりのほか，ひらがな，カタカナ，ローマ字による検索にも対応している。

見出し検索の画面では，上部に検索ボックスが用意されている。ここに検索したい特定の語を入力することで，その語の情報を獲得することができる（あるいは，品詞別タブから予めリストアップされた語を選択することでも検索を開始することができる）。参考までに，図 2 は「落とす」を検索した結果の画面を示す。「落とす」のみならず，V+V 型複合動詞の「見落とす」や「切り落とす」も含まれている。



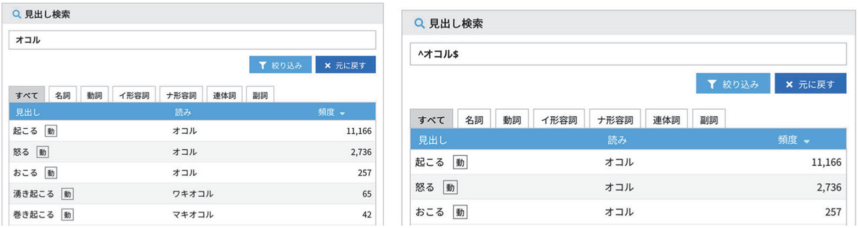
落とす

見出し	読み	頻度
落とす 動	オトス	2,209
見落とす 動	ミオトス	216
切り落とす 動	キリオトス	193
突き落とす 動	ツキオトス	86
攻め落とす 動	セメオトス	60
撃ち落とす 動	ウチオトス	45
削ぎ落とす 動	ソギオトス	33
追い落とす 動	オイオトス	26

図 2. 検索ボックスで「落とす」を指定した場合

なお，カタカナで入力し絞り込みを行うと，その読みを含むすべての見出しが表示される。例えば，カタカナで「オコル」と入力すると，図 3 左に示すように「起こる」や「怒る」などのほか，その読みを含む V+V 型複合動詞なども全て検索結果に表示される。また，図 3 右のように，V+V 型複合動詞は排除した上で「オコル」に完全一致で検索したい場合，正規表現を用いて「^オコル$」のように先頭と末尾を表す記号を入れるとよい。ひらがなで「^おこる$」と入力しても，同様の検索結果が得られる。ちなみに，「^[アイウエオ]」とすると，ア行の見出しがすべて検索できる。

図 3 の検索結果から調査したい語（今回は，「起こる」）を選択してみると，図 4 のような語彙プロファイルが一覧で表示される。この語彙プロファイルでは，左側に表示されたパタンパネル中の「名詞＋助詞」の中の「名詞＋が起こ



オコル

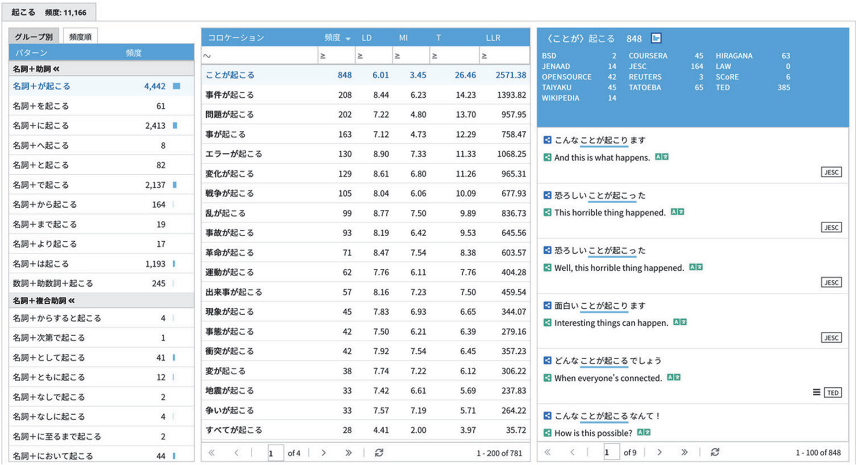
見出し	読み	頻度
起こる 動	オコル	11,166
怒る 動	オコル	2,736
おこる 動	オコル	257
湧き起こる 動	ワキオコル	65
巻き起こる 動	マキオコル	42

^オコル$

見出し	読み	頻度
起こる 動	オコル	11,166
怒る 動	オコル	2,736
おこる 動	オコル	257

図 3. 検索ボックスで「オコル」を指定した場合（左）と「^オコル$」を指定した場合（右）

る」（4,442 例）が選択されており，そのコロケーションが頻度順にリスト化され画面中央に配置されている。その中で最も頻度の高い「ことが起こる」に該当する例文が，その対訳と共に右側に表示されている。検索したい語の文法パタン，コロケーション，用例，その対訳の検索結果が瞬時に一覧表示されている点が本ツールの強みであり，現時点では唯一無二のツールであるとも言える。これは，レキシカルプロファイラー（特に，ここでは LWP）の秀逸な機能であると言ってよい。



起こる　頻度: 11,166

パターン	頻度
名詞＋助詞 «	
名詞＋が起こる	4,442
名詞＋を起こる	61
名詞＋に起こる	2,413
名詞＋へ起こる	8
名詞＋と起こる	82
名詞＋で起こる	2,137
名詞＋から起こる	164
名詞＋まで起こる	19
名詞＋より起こる	17
名詞＋は起こる	1,193
数詞＋助数詞＋起こる	245
名詞＋複合助詞 «	
名詞＋からすると起こる	4
名詞＋次第で起こる	1
名詞＋として起こる	41
名詞＋ともに起こる	12
名詞＋なしで起こる	2
名詞＋なしに起こる	4
名詞＋に至るまで起こる	2
名詞＋において起こる	44

コロケーション	頻度	LD	MI	T	LLR
~	≥	≥	≥	≥	≥
ことが起こる	848	6.01	3.45	26.46	2571.38
事件が起こる	208	8.44	6.23	14.23	1393.82
問題が起こる	202	7.22	4.80	13.70	957.95
事が起こる	163	7.12	4.73	12.29	758.47
エラーが起こる	130	8.90	7.33	11.33	1068.25
変化が起こる	129	8.61	6.80	11.26	965.31
戦争が起こる	105	8.04	6.06	10.09	677.93
乱が起こる	99	8.77	7.50	9.89	836.73
事故が起こる	93	8.19	6.42	9.53	645.56
革命が起こる	71	8.47	7.54	8.38	603.57
運動が起こる	62	7.76	6.11	7.76	404.28
出来事が起こる	57	8.16	7.23	7.50	459.54
現象が起こる	45	7.83	6.93	6.65	344.07
事態が起こる	42	7.50	6.21	6.39	279.16
衝突が起こる	42	7.92	7.54	6.45	357.23
変が起こる	38	7.74	7.22	6.12	306.22
地震が起こる	33	7.42	6.61	5.69	237.83
争いが起こる	33	7.57	7.19	5.71	264.22
すべてが起こる	28	4.41	2.00	3.97	35.72

1 - 200 of 781

〈ことが〉起こる　848

BSD	2	COURSERA	45	HIRAGANA	63
JENAAD	14	JESC	164	LAW	0
OPENSOURCE	42	REUTERS	3	SCoRE	6
TAIYAKU	45	TATOEBA	65	TED	385
WIKIPEDIA	14				

こんなことが起こります
And this is what happens.
JESC

恐ろしいことが起こった
This horrible thing happened.
JESC

恐ろしいことが起こった
Well, this horrible thing happened.
JESC

面白いことが起こります
Interesting things can happen.
JESC

どんなことが起こるでしょう
When everyone's connected.
TED

こんなことが起こるなんて！
How is this possible?

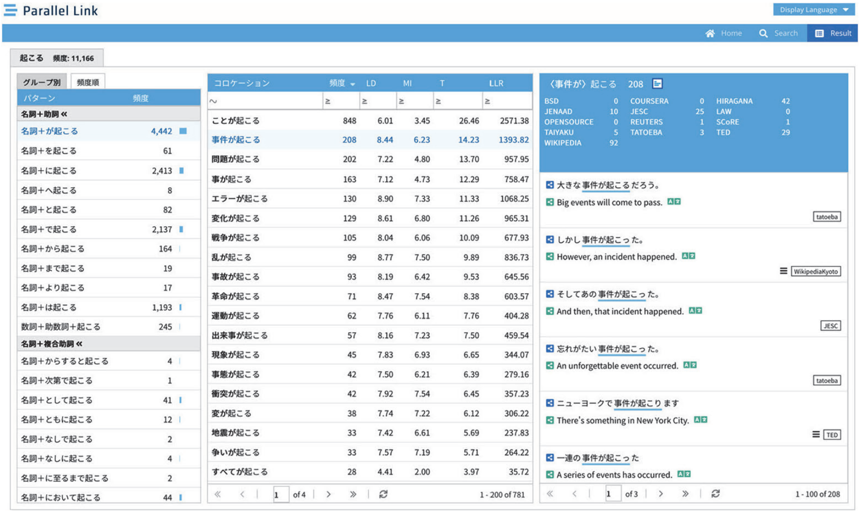
1 - 100 of 848

図 4.「起こる」の語彙プロファイル

図 4 の表示内容をもう少し詳しく解説していく。まず，左上に検索語（見出し語）の総頻度が 11,166 件であることが表示されており，その下部に「グルー

プ別」と「頻度順」のタブがある。「グループ別」には，下にスクロールしていくと「名詞＋助詞（＋起こる）」や「名詞＋複合助詞（＋起こる）」などパタンのグループ別に該当するパタンが頻度とグラフ化された割合（%）と共に表示されている。グラフにカーソルを持っていくと，実際の割合が数値で示される。「頻度順」タブをクリックすると，パタングループ別の表示から頻度順の表示に変わり，高頻度のパタンから順に頻度と割合（%）が表示される。これらの中で調べたいパタンをクリックすると，そのコロケーションリストが画面中央のパネルに表示される。つまり，検索語→パタン分析→コロケーション分析を経て該当する用例と対訳が表示される仕組みとなっており，他のコーパス検索ツールには例を見ない本ツール独自の機能であると言える。

また，図 5 では，見出し語「起こる」の中でも，「名詞＋助詞＋起こる」のパタングループのうち，「～が起こる」を指定して，さらにその中の「事件が起こる」を指定した画面である。該当する日本語の例文とその英訳が順に表示されていることが分かる。



起こる　頻度: 11,166

パターン	頻度
名詞＋助詞 «	
名詞＋が起こる	4,442
名詞＋を起こる	61
名詞＋に起こる	2,413
名詞＋へ起こる	8
名詞＋と起こる	82
名詞＋で起こる	2,137
名詞＋から起こる	164
名詞＋まで起こる	19
名詞＋より起こる	17
名詞＋は起こる	1,193
数詞＋助数詞＋起こる	245
名詞＋複合助詞 «	
名詞＋からすると起こる	4
名詞＋次第で起こる	1
名詞＋として起こる	41
名詞＋ともに起こる	12
名詞＋なしで起こる	2
名詞＋なしに起こる	4
名詞＋に至るまで起こる	2
名詞＋において起こる	44

コロケーション	頻度	LD	MI	T	LLR
～	≥	≥	≥	≥	≥
ことが起こる	848	6.01	3.45	26.46	2571.38
事件が起こる	208	8.44	6.23	14.23	1393.82
問題が起こる	202	7.22	4.80	13.70	957.95
事が起こる	163	7.12	4.73	12.29	758.47
エラーが起こる	130	8.90	7.33	11.33	1068.25
変化が起こる	129	8.61	6.80	11.26	965.31
戦争が起こる	105	8.04	6.06	10.09	677.93
乱が起こる	99	8.77	7.50	9.89	836.73
事故が起こる	93	8.19	6.42	9.53	645.56
革命が起こる	71	8.47	7.54	8.38	603.57
運動が起こる	62	7.76	6.11	7.76	404.28
出来事が起こる	57	8.16	7.23	7.50	459.54
現象が起こる	45	7.83	6.93	6.65	344.07
事態が起こる	42	7.50	6.21	6.39	279.16
衝突が起こる	42	7.92	7.54	6.45	357.23
変が起こる	38	7.74	7.22	6.12	306.22
地震が起こる	33	7.42	6.61	5.69	237.83
争いが起こる	33	7.57	7.19	5.71	264.22
すべてが起こる	28	4.41	2.00	3.97	35.72

1 - 200 of 781

〈事件が〉起こる　208

BSD	0	COURSERA	0	HIRAGANA	42
JENAAD	10	JESC	25	LAW	0
OPENSOURCE	0	REUTERS	1	SCoRE	1
TAIYAKU	5	TATOEBA	3	TED	29
WIKIPEDIA	92				

大きな事件が起こるだろう。
Big events will come to pass.
tatoeba

しかし事件が起こった。
However, an incident happened.
WikipediaKyoto

そしてあの事件が起こった。
And then, that incident happened.
JESC

忘れがたい事件が起こった。
An unforgettable event occurred.
tatoeba

ニューヨークで事件が起こります
There's something in New York City.
TED

一連の事件が起こった
A series of events has occurred.

1 - 100 of 208

図 5.「事件が起こる」を選択した場合

右側に示された各例文には，どのパラレルコーパスからの例文であるのか，出典がその右下に表示される。また，例文ウィンドウのヘッダーを見ると，「事件が起こる」の例が計 208 件抽出されたことが分かる。その下に各サブコーパ

ス別の出現回数が表示されており，その用例数を概観すると Hiragana Times が 42 例，JENAAD が 10 例，JESC が 25 例，REUTERS が 1 例，SCoRE が 1 例，TAIYAKU が 5 例，TATOEBA が 3 例，TED が 29 例，WIKIPEDIA が 92 例である一方で，他のコーパスでは全く使われていないことが分かる。各サブコーパスをクリックすれば，そのサブコーパスのみでヒットした例文だけが表示される。

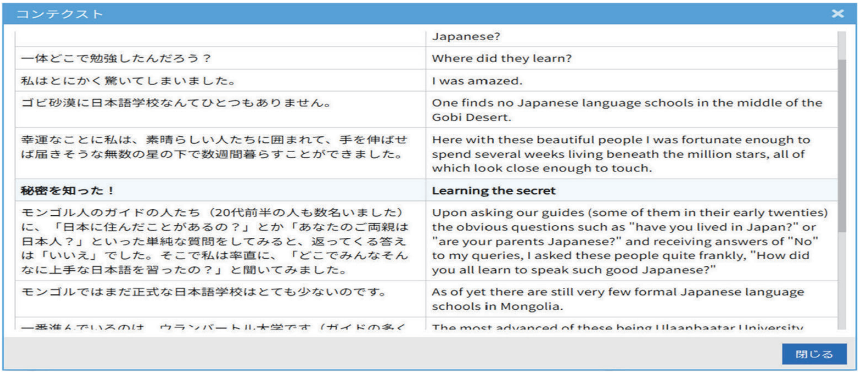
また，Ver. 1.0 の時代から拘っている点は，コロケーション抽出の際に活用できる統計指標の充実である。コロケーション情報については，NLB で搭載された頻度，MI スコア，LD スコアに加えて，T スコアと対数尤度比（LLR）の統計指標も追加している。各指標をクリックすれば，各値で並び替えが可能である。また，ヘッダーのすぐ下にあるフィルター機能を利用すれば，特定のコロケーションの検索や，任意の統計値での絞り込みができる。例えば，コロケーション列に「事件 | 事故」のように入力すると，「事件が起こる」，「事故が起こる」，「交通事故が起こる」が検索される。

2.5. Ver. 2.0 の特筆すべき機能

本節では，特に今回の Ver. 2.0 で新たに追加した 3 機能について紹介する。

2.5.1. コンテクスト（文脈）表示機能

パラレルコーパスの構造上，コンテクスト（文脈）のあるコーパス（BSD, Coursera，Hiragana Times，LAW，WIKIPEDIA）については，用例パネルでのコンテクスト表示に対応している。各用例のコーパス名の左横に（三本線）アイ



	Japanese?
一体どこで勉強したんだろう？	Where did they learn?
私はとにかく驚いてしまいました。	I was amazed.
ゴビ砂漠に日本語学校なんてひとつもありません。	One finds no Japanese language schools in the middle of the Gobi Desert.
幸運なことに私は、素晴らしい人たちに囲まれて、手を伸ばせば届きそうな無数の星の下で数週間暮らすことができました。	Here with these beautiful people I was fortunate enough to spend several weeks living beneath the million stars, all of which look close enough to touch.
秘密を知った！	**Learning the secret**
モンゴル人のガイドの人たち（20代前半の人も数名いました）に、「日本に住んだことがあるの？」とか「あなたのご両親は日本人？」といった単純な質問をしてみると、返ってくる答えは「いいえ」でした。そこで私は率直に、「どこでみんなそんなに上手な日本語を習ったの？」と聞いてみました。	Upon asking our guides (some of them in their early twenties) the obvious questions such as "have you lived in Japan?" or "are your parents Japanese?" and receiving answers of "No" to my queries, I asked these people quite frankly, "How did you all learn to speak such good Japanese?"
モンゴルではまだ正式な日本語学校はとても少ないのです。	As of yet there are still very few formal Japanese language schools in Mongolia.
一番進んでいるのは、ウランバートル大学です（ガイドの多く	The most advanced of these being Ulaanbaatar University

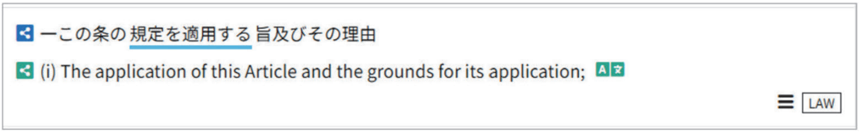
図 6.「秘密を知った！」の日英文脈状況

コンが表示されているものは，その用例の前後のコンテクストを確認できる。このアイコンをクリックすると，図 6 のように，用例（該当行）の前後 5 センテンスが表示される。

コーパスを整備する過程で重複行は削除しているが，コンテクストを表示しているコーパスについては重複行であっても前後の文脈が異なる場合があるため，重複する用例はそのまま残している。

（例）規定｜規定を適用する

LAW に次の用例が 9 件あるが，コンテクストは全て異なっている。



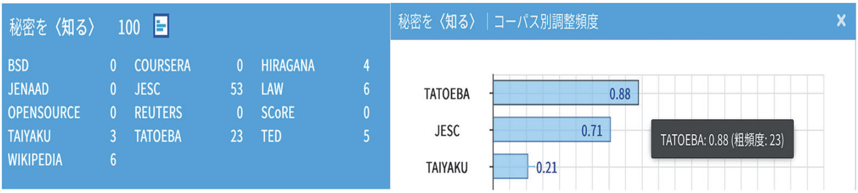
一この条の規定を適用する旨及びその理由

(i) The application of this Article and the grounds for its application;

補足情報として，コンテクスト情報のないコーパス（JENAAD, JESC, OPENSOURCE, REUTERS, TAIYAKU）は，起点言語と目標言語のテキストを文単位に分割し，目標言語から起点言語の文と類似度の高い文を見つけ出す手法で構築したものであることからコンテクストが残っていないが，これは主に著作権への配慮でもある。SCoRE や TATOEBA については文単位の例文を収集したコーパスであることから，元々コンテクストが存在していない。このような理由により，現状 BSD, Coursera, Hiragana Times, LAW, WIKIPEDIA の 5 種のみコンテクスト情報が表示される。

2.5.2. コロケーションのコーパスごとの分布グラフ表示

用例パネルのヘッダーにある▤アイコンをクリックすると，コロケーションのコーパス分布が 10 万語当たりの調整頻度で降順に表示され，棒グラフのバー



秘密を〈知る〉　100

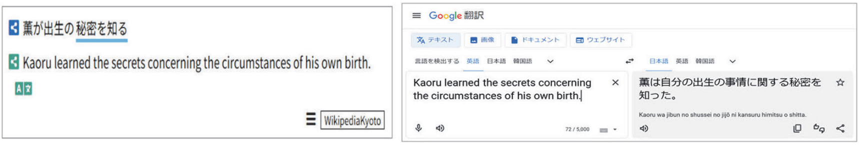
BSD	0	COURSERA	0	HIRAGANA	4
JENAAD	0	JESC	53	LAW	6
OPENSOURCE	0	REUTERS	0	SCoRE	0
TAIYAKU	3	TATOEBA	23	TED	5
WIKIPEDIA	6				

図 7.「秘密を知る」のパラレルコーパス別粗頻度表示と調整頻度グラフ機能

にマウスをかざすと調整頻度と粗頻度が表示される。

2.5.3. 音声機能と Google 翻訳リンク機能

Ver. 1.0 の時代から SCoRE コーパスから抽出された例文には音声マークが表示され，そのマークをクリックすることで音声が流れる仕組みとなっている。Ver. 2.0 では，Google 翻訳リンク機能を搭載し，SCoRE 以外の英文音声のない用例についても，Google 翻訳のサイトを開いてその音声を確認することができる。英文の後ろに表示される A文 アイコンをクリックすると，新しいタブに Google 翻訳のサイトを開き英文が表示される。音声を聞くことができる他，元の日本語と Google 翻訳の日本語訳の比較も可能である。


薫が出生の秘密を知る

Kaoru learned the secrets concerning the circumstances of his own birth.

Kaoru learned the secrets concerning the circumstances of his own birth.

薫は自分の出生の事情に関する秘密を知った。

Kaoru wa jibun no shussei no jijō ni kansuru himitsu o shitta.

図 8. Google 翻訳リンク機能

3.　Ver. 3.0 以降の改良について

3.1. 有償ライセンスコーパス「読売新聞日英文対訳コーパス」の購入検討について

今後，Ver. 3.0 以降の開発において，搭載するコーパスの拡充や検索機能の追加等をさらに進めていく。特に，学術論文や話し言葉のパラレルコーパスは欠落あるいは不足しているため，このようなジャンルの翻訳テクストを可能な範囲で増補したい。Ver. 3.0 以降の追加候補のパラレルコーパスとして，有償ライセンスで利用できる「読売新聞日英文対訳コーパス」を検討中である。現時点で 2006 年～2021 年までのデータが読売新聞東京本社メディア局に保管されており，1 年分で 34 万円と高額であるため（16 年分で 544 万円），本コーパスを完全に搭載するためには一定の研究費を獲得する必要がある。

3. 2. 検索機能の充実について

Ver. 3.0 以降では検索機能のさらなる充実も検討している。具体的には，コンコーダンサー機能や，英→日の語彙プロファイリング機能の追加を検討して

いる。レキシカルプロファイラーは，コーパス分析の初・中級者にとっては，見出し語のパタン情報やコロケーション情報を瞬時に獲得することができるため，検索結果の解釈に集中できるという利点がある。各種統計値の計算も自動化されていることから，従来の計量的な言語分析で愛用されてきたコンコーダンサーを凌駕している面もある。その一方で，レキシカルプロファイラーは予め開発者が設計した検索結果しか表示できないという制約もある。そのような制限なしに検索・分析したいコーパス分析の上級者にとっては，「特定の言語事象を対象にしたミクロな視点からの観察」を可能とするコンコーダンサーを好む者も多い。コンコーダンサーの機能も追加することで，検索語・句に関する起点言語の情報と目標言語の翻訳特性がより精緻に分析できる。

さらに，検索時の視認性を高めるために抽出された訳例中の翻訳ユニットの候補を自動的に太字で色付けする機能や，ParaConc（https://paraconc.com/）に搭載されている Hot Words 機能のように，自動的に訳語や翻訳ユニットの候補を抽出し，頻度順や統計値順に翻訳候補をリスト化する機能も搭載予定である。

また，パラレルコーパスから抽出した英文の用例を授業や教材開発，辞書の例文等で活用することを想定し，英語学習者にとって平易な英文を抽出してくれる GDEX（Good Dictionary Examples）（https://www.sketchengine.eu/guide/gdex/）の活用や，それに類似したオリジナルの機能を実装することも検討したい。特に，パラレルコーパスから抽出した英文が翻訳文である場合，信頼性 (authentic) のある英文であるとは言い難い。House（2014, p. 2）も翻訳文は “a kind of inferior substitute for the 'real thing'” と指摘する。よって，授業・教材・辞書などで英文を活用する場合には，語彙難度や文法複雑性なども調整すべきであり，その点で GDEX のような機能の活用価値は高いと言える。また，既に本ツールに実装している SCoRE（本ツールにも搭載されている教育用の平易な例文コーパス）のデータの活用も検討中である。

他にも，特定の文法パタンや共起語の翻訳を抽出した際にクリック一つで Dual KWIC 表示に切り替えできる機能や，TED コーパスについては動画ファイルへのジャンプ機能も追加したい。今後は，辞書編纂や翻訳・通訳実践（研究），対照言語学，言語教育などの分野で活用できる変則的な検索にも対応した，言語の専門家のニーズに応える翻訳コーパス集合体のオンライン検索ツール開発を目指す。

4. 想定される『パラレルリンク』の活用研究とは

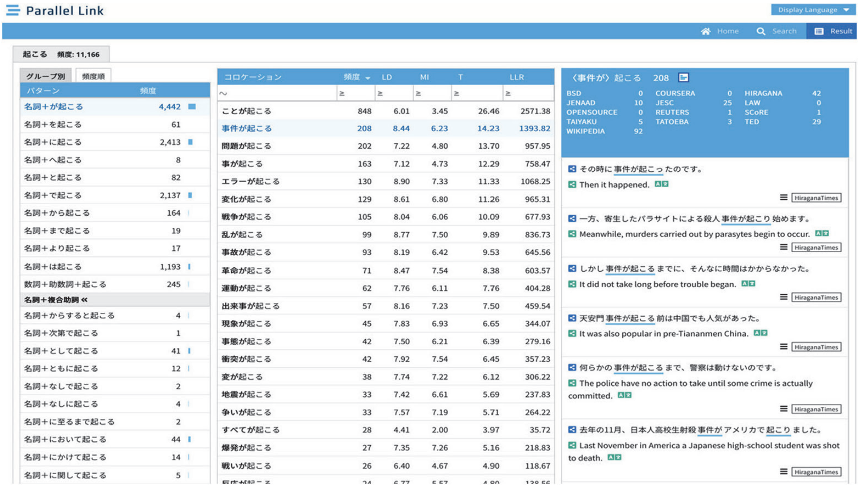
本ツールを活用した研究に，英和・和英辞書編纂時における訳語・訳例チェックのための検証的活用，翻訳･通訳実践時における現場等での参照的活用，リーディング・ライティング・翻訳の授業時など英語教育現場での実践的活用などが挙げられる。

はじめに，英和・和英辞書編纂時の検証的活用について述べる。英和・和英辞書に掲載すべき翻訳ユニットの種類・数はコーパスデータに基づく客観的指標から判断し決定すべきであると筆者らは考える。現行の英和・和英辞書に掲載されている訳語と本ツールから獲得した翻訳データとを比較検証することで，客観的指標をもって辞書記述をより信頼性あるものへと近づけることが可能となる。特に，複数のパラレルコーパスから抽出した翻訳ユニットの信頼性を量的観点から総合的にランク付けするだけでなく，ジャンルやレジスターによって訳語がどのように変化するのかを特定することで，訳語ごとにレーベル表示することが可能となり，精緻な言語事実を訳語に反映させることができる。例を挙げると, 本ツールに生起する高頻度名詞の一つに「規定」があるが, 文法パタン「複合動詞＋動詞」において「規定により＋動詞」では 13,899 例が, 「規定により〈受ける〉」は 1,303 例がヒットし，いずれも LAW コーパスからである。初めの数十例を調査すると，receive/obtain something pursuant to the provisions of〈具体的な規定内容〉といった固定化された翻訳ユニットを抽出することが出来た。よって,「規定により〈受ける〉」は法律のレーベルを貼ることができ,「受ける」は receive/obtain を用いて,「規定により」は pursuant to the provisions of〈具体的な規定内容〉といった表現を用いることが望ましいということが分かる。

また，本ツールを活用することで，現行の英和・和英辞典の記述の問題点を指摘し，仁科（2020, 2023）のようにその具体的な改善案を示すことも可能となる。具体的には，起点言語（日本語）について BCCWJ などの日本語大規模コーパスを活用した先行調査に始まり，本ツールなどの大規模パラレルコーパスによる翻訳ユニットの抽出，GDEX や SCoRE を用いて当該翻訳ユニットが含まれる簡易文の選定，その後に辞書のサンプル記述案の作成といった流れとなる。現在まで，例えば英和辞典の編纂であれば，起点言語となる英語の見出し語の選出や，その共起関係の調査にのみ単言語コーパス（ここでは英語コーパス）が活用されてきた。つまり，目標言語に関する情報，例えば二言語辞典

に掲載されている訳語については執筆者の主観によって作成されていたため，本ツールの翻訳データを活用することで，真のコーパス駆動型アプローチによる二言語辞書編纂が可能となる。

次に，翻訳・通訳実践における活用についても，やはり，一般的な辞書からは得ることの出来ない膨大な翻訳の実例を獲得出来る点に強みがあると言える。1 次翻訳は辞書などを使ってざっと粗翻訳し，2 次翻訳時に本ツールを活用して詳細にチェックする，3 次翻訳時にネイティブチェックを受けるという方法や，1 次翻訳から積極的に本ツールを活用するという方法もある。特に，現行の Ver. 2.0 では様々なジャンルの 13 種の日英・英日パラレルコーパスを搭載していることから，翻訳・通訳の用途に応じて特定のパラレルコーパスを選択し，その対訳結果を参考にする方が効率的であろう。この特定のジャンルや専門分野で好まれて使用される言語表現に注目する方法は，応用言語学における LSP（Language for Specific Purposes）の考え方に沿った TSP（Translation for Specific Purposes）であるとも言える。実際に，英訳抽出結果上部のサブコーパス別頻度の部分をクリックすると，特定のサブコーパスの用例のみが表示される。図 9 は，「起こる」→「名詞＋助詞」→「名詞＋が起こる」→「事件が起こる」の順で英訳の結果を表示させ，さらに Hiragana Times の英訳のみに絞った検索結果を示している。



起こる　頻度: 11,166

パターン	頻度
名詞＋が起こる	4,442
名詞＋を起こる	61
名詞＋に起こる	2,413
名詞＋へ起こる	8
名詞＋と起こる	82
名詞＋で起こる	2,137
名詞＋から起こる	164
名詞＋まで起こる	19
名詞＋より起こる	17
名詞＋は起こる	1,193
数詞＋助数詞＋起こる	245
名詞＋複合助詞 «	
名詞＋からすると起こる	4
名詞＋次第で起こる	1
名詞＋として起こる	41
名詞＋ともに起こる	12
名詞＋なしで起こる	2
名詞＋なしに起こる	4
名詞＋に至るまで起こる	2
名詞＋において起こる	44
名詞＋にかけて起こる	14
名詞＋に関して起こる	5

コロケーション	頻度	LD	MI	T	LLR
～	≥	≥	≥	≥	≥
ことが起こる	848	6.01	3.45	26.46	2571.38
事件が起こる	208	8.44	6.23	14.23	1393.82
問題が起こる	202	7.22	4.80	13.70	957.95
事が起こる	163	7.12	4.73	12.29	758.47
エラーが起こる	130	8.90	7.33	11.33	1068.25
変化が起こる	129	8.61	6.80	11.26	965.31
戦争が起こる	105	8.04	6.06	10.09	677.93
乱が起こる	99	8.77	7.50	9.89	836.73
事故が起こる	93	8.19	6.42	9.53	645.56
革命が起こる	71	8.47	7.54	8.38	603.57
運動が起こる	62	7.76	6.11	7.76	404.28
出来事が起こる	57	8.16	7.23	7.50	459.54
現象が起こる	45	7.83	6.93	6.65	344.07
事態が起こる	42	7.50	6.21	6.39	279.16
衝突が起こる	42	7.92	7.54	6.45	357.23
変が起こる	38	7.74	7.22	6.12	306.22
地震が起こる	33	7.42	6.61	5.69	237.83
争いが起こる	33	7.57	7.19	5.71	264.22
すべてが起こる	28	4.41	2.00	3.97	35.72
爆発が起こる	27	7.35	7.26	5.16	218.83
戦いが起こる	26	6.40	4.67	4.90	118.67
反応が起こる	24	6.77	5.57	4.80	138.56

〈事件が〉起こる　208

BSD	0	COURSERA	0	HIRAGANA	42
JENAAD	10	JESC	25	LAW	0
OPENSOURCE	0	REUTERS	1	SCoRE	1
TAIYAKU	5	TATOEBA	3	TED	29
WIKIPEDIA	92				

その時に事件が起こったのです。
Then it happened.
HiraganaTimes

一方、寄生したパラサイトによる殺人事件が起こり始めます。
Meanwhile, murders carried out by parasytes begin to occur.
HiraganaTimes

しかし事件が起こるまでに、そんなに時間はかからなかった。
It did not take long before trouble began.
HiraganaTimes

天安門事件が起こる前は中国でも人気があった。
It was also popular in pre-Tiananmen China.
HiraganaTimes

何らかの事件が起こるまで、警察は動けないのです。
The police have no action to take until some crime is actually committed.
HiraganaTimes

去年の11月、日本人高校生射殺事件がアメリカで起こりました。
Last November in America a Japanese high-school student was shot to death.
HiraganaTimes

図 9. Hiragana Times のみに出現した「事件が起こる」の日本語原文とその英訳

最後に，言語教育における活用については，特に翻訳・英語ライティングの授業などで活用が期待される。例えば，質の高い英作文を完成させるまでのプロセスで大事なことは，書き手が伝えたい内容をまずは英訳に適したやさしい日本語に一度書き換えた上で，それを英訳するという 2 段階のステップを踏む方法である。英訳に適したやさしい日本語とは，主語と述語を明確にし，長文を避け，極力標準語を使い，新語は避けるなどの配慮が施されたものであり，最近では DeepL 翻訳ツール（https://www.deepl.com/translator）など機械翻訳を有効活用する際にも留意するポイントでもある。言い換えれば，自然な日本語と自然な英語（あるいは英訳）には文化的・言語的なギャップが存在していることの裏返しでもある。これは，いわゆる Jakobson（2004; originally published in 1959）の言語内翻訳（intralingual translation）の実践でもあり，Nida and Taber（1969, p. 33）による翻訳プロセスの分析・転移・再構成を具現化した行為でもある。

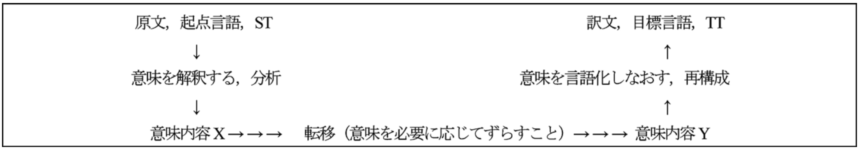

原文，起点言語，ST
↓
意味を解釈する，分析
↓
意味内容 X →→→　転移（意味を必要に応じてずらすこと）→→→ 意味内容 Y
↑
意味を言語化しなおす，再構成
↑
訳文，目標言語，TT

図 10. Nida の翻訳プロセス（Nida and Taber 1969, p. 33）

本ツールの検索から獲得できる翻訳文は，田中コーパスなどの一部を除きプロが翻訳したという意味においてオーセンティックなものばかりである。よって，どのような日本語表現において英訳との間に文化的・言語的なギャップが見られるか，例えば，日本特有の文化・言語表現が実際にどのように英訳されているかなど，予め学習者に考えさせ試訳させた上で本ツールを用いて検索させ自律学習を促すことは，翻訳上の「気づき」を促す上でも効果的であろう。模範となるプロの翻訳者が産出した翻訳に到達するためにはどのようなプロセスが必要であるのかを学習者に議論させることも，「気づき」学習を助長する上で意味があろう。そのような「気づき」に繋がるヒントがパラレルコーパスの翻訳データには埋もれている。特に，検索語の訳し方が，その共起語によって変わるという翻訳事実や，日英・英日翻訳というのはそもそも逐語的に遂行されるものではなく，時に異なる品詞に訳されたり，（あえて）省略されたりすることもある（Lost in Translation）という翻訳実態も学ぶことができる。

5. 今後の展望

以上，本稿では，最新版『パラレルリンク』Ver. 2.0のインターフェースや検索機能，想定される活用研究について紹介した。既に述べたように，搭載されている一部のパラレルコーパスの翻訳文は信頼性に欠けるため，Ver. 3.0以降では欠落・不足しているジャンルの翻訳テクストも含め，質の高い翻訳テクストを増補し，バランスの取れた日英・英日パラレルコーパス検索ツールの開発・発展を目指す。

謝辞

本研究はJSPS科研費20K00692と23K00599の助成を受けたものである。ここに，SCoREの用例コーパスを搭載することをご快諾くださった中條清美先生（元日本大学），現在SCoREの一連の研究を引き継いでおられる西垣知佳子先生（千葉大学），パラレルコーパス研究の発展を願われた故・染谷泰正先生，ならびに関係者の皆様に感謝の意を示す。

注

1. 本稿の内容は2022年10月1日にオンラインで開催された第48回英語コーパス学会および2024年10月5日〜6日に対面開催された第50回英語コーパス学会（於青山学院大学）における口頭発表の内容を大幅に修正・発展させたものである。仁科・赤瀬川（2022b），仁科（2023）にも負うところが大きい。

参考文献

House, J. (2014). *Translation: A Multidisciplinary Approach.* Hampshire: Palgrave Macmillan. doi: https://doi.org/10.1057/9781137025487

Jakobson, R. (1959/2004). On linguistic aspects of translation. In L. Venuti (Ed.), *The translation studies reader (2nd ed.)* (pp. 138–143). New York: Routledge.

Munday, J. (2016). *Introducing translation studies: Theories and applications (4th ed.)*. London and New York: Routledge. doi: https://doi.org/10.4324/9781315691862

中澤敏明・李凌寒・Matıss Rikters（2021）「ビジネスシーン対話対訳コーパスの構築と対話翻訳の課題」『言語処理学会第27回年次大会発表論文集』1375–1380.

Nida, E.A., & Taber, C.R. (1969). *The theory and practice of translation.* Leiden: E.J. Brill.

仁科恭徳（2014）「実践で学ぶコーパス活用術：第11回パラレルコーパスの可能性」『研究社WEBマガジンLIngua（リンガ）』オンライン.

仁科恭徳（2020）「日英パラレルコーパス WikipediaKyoto-LWP を用いた和英辞典の記述改善案について―「X を固める」の場合―」『英語コーパス研究』第 27 号：1–21.

仁科恭徳（2023）『パラレルコーパス言語学の諸相』東京：開拓社.

仁科恭徳・赤瀬川史朗（2021）「日英・英日パラレルコーパスオンライン検索ツール『（仮称）パラレルリンク』（Ver. 1.0）の開発に向けて（中間報告）」『英語コーパス学会大会予稿集 2021』25–30.

仁科恭徳・赤瀬川史朗（2022a）「『パラレルリンク』（Ver. 1.0）の開発―パラレルコーパス研究の概観とコーパス整備―」『英語コーパス研究』第 29 号：63–78.

仁科恭徳・赤瀬川史朗（2022b）「日英・英日パラレルコーパス検索ツール『パラレルリンク』（Ver. 1.20）―インターフェース，検索機能，活用研究などについて―」『英語コーパス学会大会予稿集 2022』7–12.

パルデシプラシャント・赤瀬川史朗（2011）「BCCWJ を活用した基本動詞ハンドブック作成－コーパスブラウジングシステム NINJAL―LWP の特長と機能―」『現代日本語書き言葉均衡コーパス完成記念講演会予稿集』205–216. 東京：国立国語研究所.

染谷泰正・赤瀬川史朗・山岡洋一（2011）「大規模翻訳コーパスの構築とその研究および教育上の可能性」『日本メディア英語学会第 1 回年次大会発表資料』1–15.

（仁科　恭徳　神戸学院大学）

（赤瀬川史朗　Lago NLP）

「研究ノート」

中・高生向け DDL 英文法学習支援サイトの開発
—DDL 普及の障壁解消の取り組み—

西垣知佳子・赤瀬川史朗

Abstract

This study explores the development and implementation of data-driven learning (DDL) tools for middle and high school students to overcome barriers in English language education. DDL, proposed by Johns (1991), emphasizes the use of corpus tools for pedagogical purposes, enabling learners to discover language patterns through active analysis of linguistic data. Despite its potential, DDL has not been widely adopted in primary and secondary education due to challenges such as the complexity of authentic texts and the need for specialized training. To address these issues, we developed hDDL2, a user-friendly DDL support software tailored for younger learners. The hDDL2 tool incorporates an educational corpus adjusted for vocabulary and grammar levels appropriate for junior and high school students. hDDL2 effectively supports inductive, student-centered learning, aligning with modern educational goals. The tool's features, including easy-to-use search functions and grammar quizzes, facilitate the discovery and retention of language rules. This paper discusses the progressive development of hDDL2, and its application to beginner-level classroom settings. The results suggest that hDDL2 can bridge the gap between theoretical DDL approaches and practical classroom implementation, promoting more engaging language learning experiences for beginning-level students.

1. はじめに

Data-driven learning（DDL）は，Johns（1991）によって提唱された英語学習法であり，“the use of corpus tools and techniques for pedagogical purposes in a foreign/second language”（Boulton & Vyatkina, 2021, p. 68）と定義される。Johns

によると, DDL における学習者のタスクは “to ‘discover’ the foreign language”（p. 1）であり，教師は学習者が “learn how to learn”（p. 1）することを支援する。すなわち, DDL を通して, 学習者は言語データ（コーパス）を主体的に観察し，自分の力で言語データを分析して語彙や文法の規則を発見する力を身に付ける。さらに，DDL は学習者が協働して言語データを分析して知識を構築するプロセスを支援する（Corino & Onesti, 2019）。こうした学習者主導型の学習スタイルは，生徒が知識と技能を活用して，課題を発見して解決するために必要な思考力，判断力，表現力の育成を目指す現代の学習指導要領の考え方（文部科学省，2017，2018）と合致する。一方，DDL は，提唱されてから 30 年以上経過するものの，期待されたほど普及していない。そこで，本研究では，現代の英語教育が求める学び方と符合する DDL を，学校教育に導入し普及させる具体的な方法として，DDL 英文法学習支援サイトを提案する。そして，本稿の最後の章では，中学 2 年生（130 名）を対象に実施した意識調査の結果を踏まえて，DDL 英文法学習支援サイトの開発目的が達成されているかどうかを検討する。

2. DDL に関する先行研究調査

本章では，これまでに行われた DDL 普及に関わる先行研究について述べる。

2.1 コミュニケーション重視の指導と文法指導

DDL の学び方は帰納的であるが，コーパスを利用して語彙と文法の明示的知識を身に付けることを目的とすることから，「形式偏重型」の学習方法に分類される（久保田，2018）。一方，今日の外国語／第二言語教育では，コミュニケーションが適切になされていれば，第二言語の文法の知識は自然に獲得されるという「意味偏重型」の暗示的学習が推奨されている。本節では，現代の外国語／第二言語教育の状況において，文法指導がどのように扱われているかについて検討する。

例えば Nation（2024）は，“Learning can occur without teaching, and most successful language learning occurs without teaching.”（p. 204）と述べ，語彙・文法を暗示的に指導するように推奨している。さらには，“teaching which focuses on the deliberate learning of items should make up only a very small part of a language course.”（p. 204）とも述べ，意図的な学習はできるだけ少なくするように注意

を与えている。

その一方で，Loewen（2020）は，第二言語の習得について，“interaction may result in a better ability to communicate, interaction without any attention to language forms does not necessarily improve linguistic accuracy.”（p. 65）と述べて，明示的な指導なくしては言語の正確さは身に付かないことを指摘している。さらに，白畑（2015）は，日本の学習環境に照らして，「日本のような英語の習得に不利な言語環境では，明示的な指導が効果的である」と言う。そして，「母語の知識，また，自分の考えを母語で言語化できる認知能力と分析能力の利用が有効」であると指摘している。

以上のように，コミュニケーション重視の現代の英語教育の流れの中にあっても，文法指導は不可欠であり，重要であると言う考え方は根強い。

2.2 文法学習の重要性

英語学習において，文法指導が重要であることは，中学生の文法力を調査したテストの結果から示唆される。例えば，尼寺・西垣（2024）は，公立中学校の3年生89名を対象に，高校入試直前の12月に，be動詞と一般動詞の使い分けの理解度を調査した。具体的には，「あなたはテレビを見ますか。」という日本語の文を提示し，これを英訳させた。この設問ではbe動詞と一般動詞の区別が出題意図であったため，スペルミスや冠詞の誤りは採点の対象としなかった。その結果，“Are you watch TV?” や “Do you watching TV?” 等の誤答があり，正答率は7割だった。約3割の生徒が，be動詞と一般動詞の違いを理解しないまま，中学校を卒業していく状況であった。

また，金谷ほか（2017）では，全国の上位校の高校1年生224名，中位校の高校2・3年生555名，下位校の高校1年95名（合計974名）が，「私の友達はいつも優しいです。」を “My friends are always kind.” に,「あの赤いペンは, キッチンにありました。」を “That red pen was in the kitchen.” に英訳する和文英訳問題10問（10点満点）に回答した。文法的に正しく，日本語の文意を正確に伝えている解答は正解とし，スペルミスは減点の対象外とした。その結果，満点を獲得したのは上位校の2名のみであったと報告している。

さらに，Kakiba and Nishigaki（2021）は，中学1年生の2学期に，教科書の学習ターゲットである “This is...,” “That is...” を音声英語を通して「導入」した後,「これはあなたの本です。」「これはケンの本です。」等，学習ターゲットを含む5問の和文英訳課題を「導入」の前後で実施した。その結果，“*This is you're book.” や “*That is you are book.” と書いた生徒が多数いたことを報告している。小学生 3 年生から慣れ親しみ，中学入学後まもなく，明示的に学んだ “your” を，生徒は “you’re” や “you are” という形で内在化していたことが見て取れた。また，“Kens” のように所有格におけるアポストロフィの欠落や，音声英語では大文字と小文字の区別がないために，名前を小文字で書き始めるという誤りも目立っていた。

以上の学習者の実態から，コミュニケーション重視の指導においても，文法に意識を向ける明示的な指導が重要であることは明らかであろう。

2.3 DDL の実践事例

前節で示唆されたように, 学習者の文法学習における課題が明らかになる中，効果的な文法指導法として近年関心を集めているのが DDL である。DDL は，Boulton and Cobb（2017）によるメタ分析の結果，英語学習方法として指導効果が高いとされている。また，主に日本人学習者を対象としたメタ分析の結果からも DDL の指導効果が報告されている（Mizumoto & Chujo, 2015）。しかし，従来の DDL は，大学生を対象としている指導が多数を占めており，大学生以下の小中高生を対象とした DDL は少ない（Crosthwaite，2020）。そして，DDL 研究者は DDL を広く言語教育に導入しようとしているものの，DDL は大学の環境に限定されているという指摘もある（Boulton, 2010）。Chambers（2019）は，94% 以上の DDL 研究において，コーパス言語学の専門家が高等教育の上級レベルの学習者を対象に DDL を導入していると報告している。

日本国内での DDL 実践研究を見ると, 初等教育における DDL の実践として，Takahashi and Fujiwara（2016）がある。日本の公立小学 6 年生 52 名に対して，名詞の単数形と複数形の指導にペーパー版 DDL を活用した。学習内容に関する音声テストの結果，DDL 群では，事前テストと直後の事後テストで有意な得点の向上が見られ，指導 17 日後の遅延テストにおいても得点下降は見られず定着が良かった。一方，従来型群では得点の向上は見られなかった。

中学生に DDL を適用した例としては，Nishigaki and Kakiba（2023）がある。中学 1 年生 57 名が動詞の 3 人称単数現在形を学習した。ペーパー版 DDL を通して，生徒の文法知識がどのように向上するか，また，生徒はどのような発見をして，どのようにしてその発見に至ったのか等について調査した。DDL クラスでは，生徒が個々にルールを発見した後，発見をグループで共有し，最後に教師が生徒の発見をまとめた。従来の教師主導のクラスでは，生徒は教師

から説明を受けて練習問題を解いた。両クラスともに指導 1 週間後の事後テストで得点は上昇したが，DDL クラスはクラス全体に得点上昇があり，下位群の得点が上位群に近づいた。一方，教師主導のクラスでは，得点の分布に上位群と下位群で 2 つの山が生じた。

以上のように，日本国内で，小学校，中学・高校での DDL の活用が広がってきているものの，DDL 実践研究の報告数は限られている。

2.4 DDL の普及を妨げる要因

DDL が提唱されてから 30 年以上経過しているが，当初期待されたほどの広がりを見せていない。そこには，DDL の普及を阻害する要因がいくつかあると考えられる。まず，DDL が扱う *authentic* な英語データが初級者には難しすぎるということがある（Allan, 2009）。そのため，初級者の指導には，初級学習者向けに作られた教育用コーパスが必要である（Chujo, Oghigian, & Akasegawa, 2015）。また，DDL では，検索方法や発見の方法に慣れる必要がある。Boulton and Cobb（2017）のメタ分析や，初等・中等教育での DDL を論じている Crosthwaite（2020）によると，DDL の活用には検索ツールの操作に慣れることが必要であり，加えて DDL の学びを効果的にするには指導者からのガイダンスが必要であることも強調されている。

2.5 DDL の普及に向けた取り組みと研究目的・研究課題

学習指導要領（文部科学省，2017，2018）に照らし合わせて，筆者らは，帰納的で自発的で，かつ生徒中心の学びを実現する DDL を中学・高校の英語授業に取り入れたいと考えた。そのために，DDL が抱える課題を解決するべく，hDDL と呼ばれる中・高生向けの DDL 英文法学習支援サイトを開発した。hDDL が搭載するコーパスは，日本の初級大学生向けに開発された SCoRE（Chujo, Oghigian, & Akasegawa, 2015）にならって，語彙･文法レベル，文の長さ，文の内容を中学・高校レベル用に調整した教育用英語例文を搭載している。また，hDDL は操作性が高く，調べたい文法項目のボタンをクリックするだけで検索できる。さらに，無料で利用でき，登録の必要がないので，学校でも自宅でも自由に利用できる。hDDL は，2019 年から更新を重ねており，Version 1.20 にあたる hDDL（1.20）については，Nishigaki and Akasegawa（2023）で報告した。こうして hDDL は, DDL が抱える課題を解決してきた。しかし, hDDL（1.20）を運用していると，さらなる改善点や機能追加の要望が出てきた。そこで，そ

れらに応えるためにhDDL（1.20）を大幅に改修し，Version 2.00となるhDDL（2.00）（以下，hDDL2）を開発することを目的として本研究に取り組んだ。設定した研究課題は，以下の2点である。

研究課題1　hDDL2が搭載するコーパスは，日本の中・高生向けの教育用英語例文コーパスとして適切なレベルであるか。

研究課題2　hDDL2は，中・高生向けのDDL英文法学習支援サイトとして，入門期の学習者にも操作方法は簡単であるか。

なお，このhDDLシリーズの開発にあたっては，開発計画の立案，コーパスの作成，DDL教材作成，DDL指導実践と効果検証，教師へのDDL研修は第一筆者が行っている。また，システム開発全般を第二筆者が担い，教師と生徒の双方に使いやすいhDDLシリーズを作成している。

3. hDDL2の概要と機能

本章では，生徒の学習と教師の指導を支援するために，情報過多を抑制し，生徒の発見を促す機能の強化を目的として，大幅に改修したhDDL2について述べる。

3.1 hDDL2の教育用コーパス

hDDLが搭載する教育用コーパスは，扱う文法項目と英語例文の追加および修正を定期的に行っており, hDDL2では, 20の文法項目の学習が可能である（表1）。hDDL2が搭載する英文は，2024年11月現在，計5,370文で，その文の種類の内訳は表2のとおりである。最短の文は1語，最長の文は19語，英文1文の長さ（1文に含まれる単語の数）の平均は6.12語である。一方，現行の中学校英語検定教科書の1文の平均の長さは9.34語であることから，英文の長さの観点から，hDDL2は生徒のレベルに合致している。また，英文の長さは英文の複雑さと連動している（Yano, Long, & Ross, 1994）ことから，文構造の観点からも，hDDL2の英文は中・高生のレベルに合致していると推察される。

次に，New Word Level Checker（Mizumoto, 2021）により，新JACET8000を基準にしてhDDL2の語彙レベルを分析した結果，固有名詞が5.91%含まれており，L2レベル（2,000語）までの語彙で95.95%がカバーされている。学習

表 1. hDDL2 が扱う 20 の文法項目

1 命令文	6 存在構文	11 動名詞	16 完了形
2 be 動詞	7 未来の文	12 比較	17 関係代名詞
3 一般動詞	8 助動詞	13 間接疑問文	18 関係副詞
4 進行形	9 文のかたち	14 受動態	19 仮定法
5 過去形	10 不定詞	15 分詞	20 接続詞

表 2. hDDL2 が搭載する英文の種類と例文数

文の種類	肯定文	疑問文	否定文	命令文	総数
文の数	3,244	984	837	305	5,370

指導要領（文部科学省，2017）では，小学校と中学校を合わせて 2,200 語から 2,500 語を学習することになっているため，語彙の観点からも中・高生レベルに合致している。

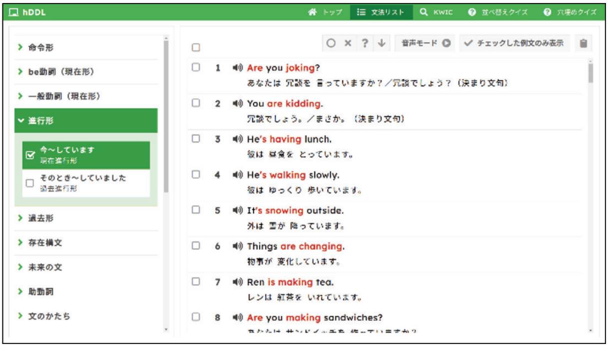
以上の分析結果から，hDDL2 は，中学・高校レベルに合致したコーパスを備えていると言えると考えられ，研究課題 1 は妥当なレベルであったと言えよう。

3.2 hDDL2 を構成するツール

hDDL は中・高生にとって使いやすいツールを目指し，1 年に 1 回から 2 回のペースで定期的に改修を行い，hDDL2 は 5 回目の改修版である。hDDL2 は，(1) 文法リスト，(2) KWIC，(3) 並べ替えクイズ，(4) 穴埋めクイズという 4 つのツールで構成されている。このうち，(1) と (2) は「文法学習ツール」で，(3) と (4) は学習した文法事項の理解度を確認する「評価ツール」である。hDDL2 の画面右上にある 4 つのいずれかのツール選択ボタンを押すと，そのツールの画面が表示される（図 1）。

図 1. ツール選択ボタン

以下に，hDDL2 の各ツールの概要を述べる。1 つ目は，「文法リスト」であ



□

□ 1 Are you joking?
あなたは 冗談を 言っていますか？／冗談でしょう？（決まり文句）

□ 2 You are kidding.
冗談でしょう。／まさか。（決まり文句）

□ 3 He's having lunch.
彼は 昼食を とっています。

□ 4 He's walking slowly.
彼は ゆっくり 歩いています。

□ 5 It's snowing outside.
外は 雪が 降っています。

□ 6 Things are changing.
物事が 変化しています。

□ 7 Ren is making tea.
レンは 紅茶を いれています。

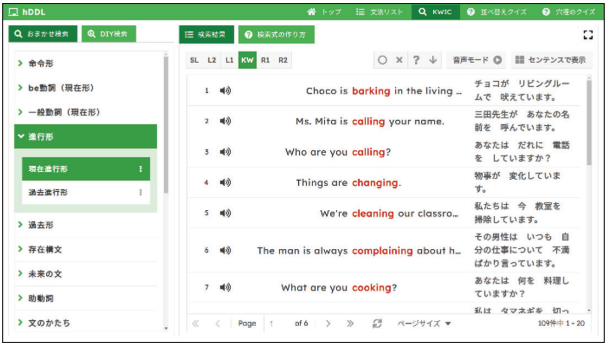
□ 8 Are you making sandwiches?

図 2. 文法リストの画面例

る（図 2）。画面左側に文法項目のリストが表示される。学習したい文法項目をクリックすると，その下にサブ項目が現れる。図 2 の画面のように，「進行形」をクリックし，次にサブ項目である「現在進行形」をクリックすると，右側の用例パネルに，現在進行形の文構造を含む例文が表示される。その中で，現在進行形の文の枠組みである「be 動詞＋動詞の -ing 形」の部分が赤色で示される。英文の左側のスピーカーアイコンをクリックすると，その例文の発音が確認することも可能である。

hDDL シリーズの特長として，英文に加えて日本語訳が併記されていることが挙げられる。日本語訳は，英文と対応させて文の構造を理解しやすいように直訳調を採用している。日本語訳があることは，英文の意味の理解だけでなく，未知語の意味の確認にも利用できることから，発見学習を支援するとともに，日本語と英語の文構造の違いを理解する手助けをしたり，DDL の難易度を下げたりすることにも役立っている。

2 つ目のツールは「KWIC」（Key Word in Context）である。KWIC では，キーワードが画面右の用例パネルの中央に縦に並ぶように用例を配置する。現在進行形の場合，「動詞の -ing 形」をキーワードとして中央縦一列に英文が並ぶ（図 3）。hDDL2 における KWIC の最大の特長は，検索式なしで KWIC を表示できる点である。hDDL2 ではこの検索方式を「おまかせ検索」と呼んでいる。生徒は調べたい文法項目を選んでクリックするだけで，KWIC を表示することができる。通常の KWIC 検索では，検索式の入力に時間をとられてしまうことが多いが，おまかせ検索によって，生徒は検索式の入力に煩わされることなく，すぐに英文の観察を始めることができる。



1	Choco is barking in the living ...	チョコが　リビングルームで　吠えています。
2	Ms. Mita is calling your name.	三田先生が　あなたの名前を　呼んでいます。
3	Who are you calling?	あなたは　だれに　電話を　していますか？
4	Things are changing.	物事が　変化しています。
5	We're cleaning our classro...	私たちは　今　教室を掃除しています。
6	The man is always complaining about h...	その男性は　いつも　自分の仕事について　不満ばかり言っています。
7	What are you cooking?	あなたは　何を　料理していますか？

図 3. KWIC の画面例

画面右側の用例パネルの左上にある［SL］から［R2］までの 6 つのボタンは並べ替えのためのボタンである。［L2］，［L1］はキーワードの左 2 語目，左 1 語目でそれぞれ並べ替える。［KW］はキーワードで並べ替える。［R1］，［R2］はキーワードの右 1 語目，右 2 語目で並べ替える。［SL］（Sentence Length）は，通常のコンコーダンサには見られないセンテンスの語数による並べ替えである。［SL］で並べ替えを行うと，センテンスの短いものから順に英文が表示される。短い例文ほどルールが発見しやすいので，生徒は画面の上に表示される短い英文から順に観察すればよい。

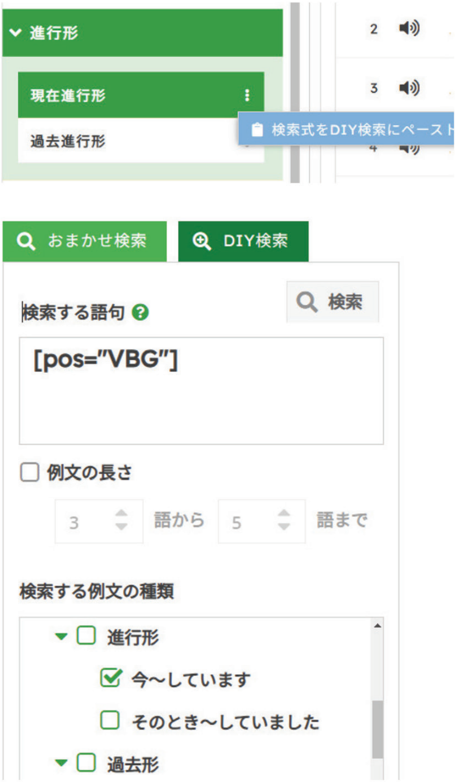
なお，KWIC ツールには，おまかせ検索の他に「DIY 検索」がある。DIY 検索は，入力した検索式にしたがって検索結果を表示する通常の KWIC 検索に相当する機能である。DIY 検索では，コーパスクエリ言語（corpus query language）を利用した本格的な検索ができるが，生徒が最初からそのような高度な検索を行うことは難しいので，hDDL2 では，おまかせ検索から DIY 検索への橋渡し的な機能を用意している。

まず，おまかせ検索で文法項目とサブ項目を選ぶと，用例パネルに英文が KWIC 表示される。このとき，サブ項目の右にある 3 点のアイコン（⋮）をクリックすると，［検索式を DIY 検索にペースト］というメニューが表示される。この項目をクリックすると，おまかせ検索であらかじめ設定されている検索式と，検索対象となる文法項目が DIY 検索の画面に自動的にセットされる。図 4 の下の部分を見ると，［検索する語句］に現在進行形の検索式［pos=“VBG”］がセットされ，［検索する例文の種類］で［今～しています］が選択されているのがわかる。この設定を利用して，生徒は自由に検索範囲を広げたり狭めたりして，

自律的な発見学習につなげていくことができる。筆者が参観する授業では，教科書に出現する単語の使い方を hDDL2 で確認したり，英作文の時に辞書の代わりに DIY 検索をしたりする生徒がいる。現在，多くの学校で PC は文房具のように身近な道具になっており，生徒は，使い方が簡単な hDDL2 を手軽に利用できるようになっている。

2 つの文法学習ツールの次に，2 種類の「評価ツール」を紹介する。これらは，「文法リスト」と「KWIC」で学習した文法知識の理解度を確認するアウトプットツールである。まず，1 つ目は「並べ替えクイズ」である。日本語訳を手がかりにして，与えられた語句を並べ替えて英文を完成させる。

2 つ目は hDDL2 に新しく搭載された評価ツールの「穴埋めクイズ」である。日本語訳をヒントにして，空欄の枠に入る単語のスペルを正しくキーボードから入力して，英文を完成させる（図 5）。現在の英語の授業では，かつてほどスペリングの知識は重視されていないが，定期試験や入試で


進行形

現在進行形

過去進行形

2

3

検索式をDIY検索にペースト

おまかせ検索

DIY検索

検索する語句

検索

[pos="VBG"]

□ 例文の長さ

3 語から 5 語まで

検索する例文の種類

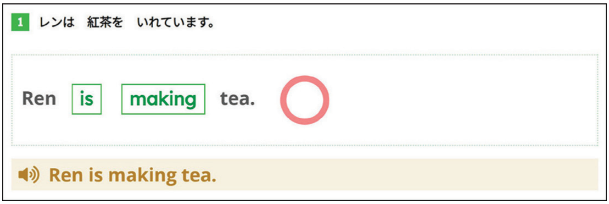
□ 進行形

☑ 今〜しています

□ そのとき〜していました

□ 過去形

図 4. おまかせ検索と DIY 検索


1 レンは 紅茶を いれています。

Ren is making tea. ○

Ren is making tea.

図 5. 穴埋めクイズの画面例

は正しいスペルが求められるので，単語を入力する穴埋めクイズは生徒から喜ばれている。定期試験前になると，生徒はこの 2 種類の評価ツールを試験準備に利用している。

3.3 英文ルールを発見しやすくする機能

hDDL2 では，生徒が言語規則を発見しやすくなるように，さまざまな工夫をしている。ここでは，そのような機能のいくつかを紹介する。

1）文の種類によるフィルタリング

通常の検索では，肯定文，疑問文，否定文，命令文の 4 種類の例文が混在して画面に表示される。多様な種類の文が同時に並ぶと，どの文のどの部分を見て，何を発見すればよいのか焦点を絞りにくい。そこで，hDDL2 の文法リストと KWIC の文法学習ツールでは, 用例パネルの上部にある［○ × ？ ⬇］（左から順に「肯定文」「否定文」「疑問文」「命令文」を表す）のボタンを押すと，文の種類を絞り込む（フィルタリングする）ことができる。

2）2 つの文法項目の比較

文法学習の際，既習と新出の文法項目を並べて比較することで，両者の違いを際立たせ，新たな知識体系を築くことがある。hDDL2 では，例えば，［過去進行形］を表示している状態で，［現在進行形］をクリックすると，用例パネルの画面上下に過去進行形の英文と現在進行形の英文を同時に表示できる。この機能を利用すると，新たに学ぶ過去進行形を理解する際に、既に学習した現在進行形と比較しながら, 両者の文法規則の違いを意識的に学ぶことができる。

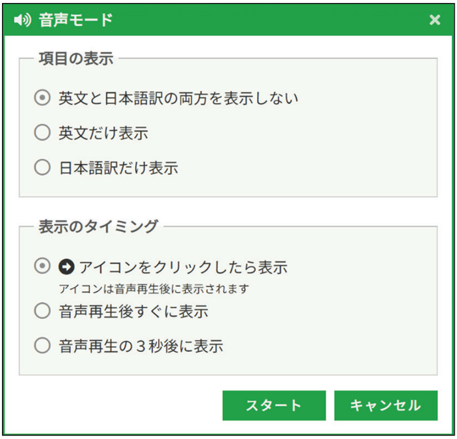
3）音声モード

hDDL2 では，英文や日本語訳を隠して，英文の音声のみを聞いて，その英文に共通する文法ルールを発見するという音声準拠 DDL を追加した。学習ツールの上部にある［音声モード］ボタンを押すと図 6 のような音声モード設定画面が表示される。音声を聞くときに，英文と日本語訳のいずれかまたは両方を非表示にするか，また，再表示するタイミングをどうするかを細かく設定できる。英文や日本語訳を表示するタイミングはリスニング学習において重要で，音声を聞いた後に一定の間を置くことで、音声から英文を想起する時間を確保できる。

4）ページサイズの変更

KWIC では，検索結果をページごとに表示する。初期設定では，1 ページに 20 の例文が表示されるが, この機能により 1 ページに表示する例文の数を 3 文,

5 文，10 文，20 文，50 文，100 文，500 文から選ぶことができる（図 7）。1 画面に表示する例文を，3 文または 5 文と数を少なくすると，焦点が絞りやすくなり，限られた例文をじっくり観察できるので発見につながりやすい。特に初級学習者にとって有効な機能である。一方，100 文や 500 文のような大きなページサイズに変更すると，ページの切り替えが不要となり，全体を俯瞰する際に便利である。



音声モード

項目の表示

◉ 英文と日本語訳の両方を表示しない

○ 英文だけ表示

○ 日本語訳だけ表示

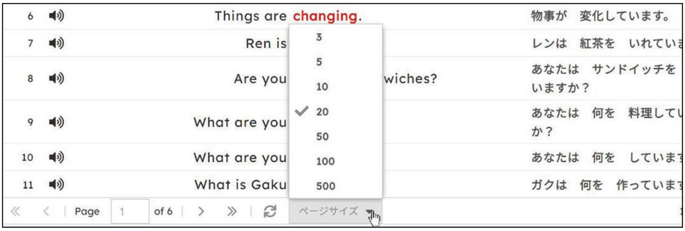
表示のタイミング

◉ アイコンをクリックしたら表示
アイコンは音声再生後に表示されます

○ 音声再生後すぐに表示

○ 音声再生の３秒後に表示

図 6. 音声モードの設定画面



6 Things are changing.	物事が　変化しています。
7 Ren is	レンは　紅茶を　いれてい
8 Are you　wiches?	あなたは　サンドイッチを いますか？
9 What are you	あなたは　何を　料理して か？
10 What are you	あなたは　何を　していま
11 What is Gaku	ガクは　何を　作っていま

図 7. ページサイズの変更画面の例

4. hDDL2 の使用に関する意識調査

本章では，実際に生徒が DDL を利用して英文法を学習した際に，hDDL2 の使い方の難しさの程度について，「簡単だ」と感じるかどうか，また，「hDDL2 が英文法学習ツールとして役立つ」と感じるかどうか等，hDDL2 の活用に関する意識調査を実施した。

4.1 調査の方法

参加者：千葉県内の中学生 2 年生 130 名。この中学校では，入学時に，保護

者から，教育研究参加の同意を得ている。

実施時期：hDDL2 を利用開始時に，当該の英文法学習支援ツールの操作を容易に感じるかどうかについて，質問紙調査を実施した。質問項目の内容を表 3 に示す。参加者は 5 件法（5 そう思う，4 少しそう思う，3 どちらでもない，2 あまりそう思わない，1 そう思わない）により回答した。

4.2 調査結果

質問項目とそれに対する生徒の回答の平均値を表 3 に示す。まず，Q1 の回答の平均値は 4.08 であり，生徒は hDDL2 の使い方を容易だと感じていることがわかる。この結果，研究課題 2 は支持されたと考えられる。また，Q3 の平均値は 4.17 であり，生徒は，hDDL2 は文法学習に役立つと感じていることがわかる。このことから，hDDL2 は，文法学習支援サイトという学習目的を達成していると考えられる。その一方で，Q2 の「授業以外でも hDDL2 を使いたい」という項目の平均値は 3.41 と，他の回答と比べて平均値は高くない。その理由としては，DDL の使用を始めたばかりであり，授業以外で一人で使いこなせるようになるには時間がかかることが推測できる。また, Q4 の結果から，文法学習の方法として，「先生の説明を聞いて学習する受動的方法」を好む学習者が少なくないことが反映されていることが考えられる。文法学習に対する志向は生徒によって異なるため，hDDL2 を，個別最適な学びを考える際の一つの選択肢として位置付けることができる。また, これまで, hDDL2 の改修は，授業利用の支援を目的として考えてきたが，今後は，個別最適な学びやホームスクーリングへの対応を視野に入れ，個別学習を支援する機能のさらなる充実を図っていきたい。

表 3. 質問項目とそれに対する生徒の回答の平均値

Q 1	hDDL2 の使い方は簡単だ	4.08
Q 2	hDDL2 を英語の授業の時間以外でも使いたい	3.41
Q 3	hDDL2 は英語の「文法の学習」に役立つ	4.17
Q 4	学校の授業は，先生の説明を聞いて学習する方法が好きだ	3.85
Q 5	学校の授業は，自分ひとりで考えて学習する方法が好きだ	3.17
Q 6	学校の授業は，友だちと意見を交換して学ぶのが好きだ	3.92

5. 結果とまとめ

本稿では，文法学習の重要性やDDLの実践事例を踏まえた上で，まず，DDL普及を阻害する障壁を検討し，DDLが取り組むべき課題として，学習者の英語力に対してコーパスの英文の難易度の高さ，検索ソフトの操作の煩雑さ，生徒を発見に導く支援の必要性という点があることを明らかにした。続いて，これらの課題を解決することを目指して開発したhDDLについて，開発のプロセスと機能を紹介し，我々が長年取り組んできた成果である最新のhDDL2を紹介した。hDDL2は，搭載する教育用コーパスがターゲットとする学習者に適切なレベルであること（RQ 1），また，入門期の生徒が利用した際に，使い方が簡単であると評価されたこと（RQ 2）が確認でき，本研究における2つの研究課題は承認されたと考える。

hDDL2の開発によって，DDL普及を阻む障壁のいくつかを解決したが，全てを解決できたわけではない。ホームスクーリングにも利用できる，個々の学習者のニーズに配慮した機能や学習ガイド，また，DDL授業者への支援の充実など，解決すべき課題は残されている。今後とも，日本の英語教育におけるDDL普及に向けての取り組みを継続し，DDLを通して探究的な英文法の学びのスタイルを学校教育の現場に紹介する取り組みを続け，意味重視の指導に知識の習得を統合する授業設計を検討していきたい。

謝辞

hDDLシリーズの開発は，16H03441（基盤B），20H01277（基盤B），24K00083（基盤B）（いずれも研究代表者西垣知佳子）の支援を受けて行われました。

参考文献

Allan, R. (2009). Can a graded reader corpus provide 'authentic' input? *ELT Journal, 63*(1), 23–32.

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

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, *67*(2), 348–393.

Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning & Technology, 25*(3), 66–89.

Chambers, A. (2019) Towards the Corpus revolution? Bridging the research–practice gap, Lang. Teach. 52 (4) (2019) 460–475,

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). John Benjamins.

Corino, E., & Onesti, C. (2019). Data-driven learning: A scaffolding methodology for CLIL and LSP teaching and learning. *Frontiers in Education, 4*, Article 16.

Crosthwaite, P. (Ed.). (2020). *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge.

Johns, T. (1991). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Johns & P. King (Eds.), *Classroom concordancing. English Language Research Journal, 4*, 27–45.

Kakiba, A., Nishigaki, C., & Oghigian, C. (2021). Implementation of Data-Driven Learning (DDL) in English classes for first-year junior high school students. *Bulletin of the Faculty of Education, Chiba University*, *69*, 179–187.

金谷憲・臼倉美里・大田悦子・鈴木祐一・隅田朗彦.（2017).『高校生は中学英語を使いこなせるか ?』. アルク.

久保田章（2018). 文法の学習と指導. 望月昭彦･久保田章･磐崎弘貞･卯城祐司（編),『新学習指導要領にもとづく英語科教育法』(第 3 版, pp. 226–240). 大修館書店.

Loewen, S. (2020). *Introduction to instructed second language acquisition* (2nd ed.). Routledge.

Mizumoto, A. (2021). *New Word Level Checker*［Web application］.

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

尼寺圭悟･西垣知佳子（2024)「英語の基礎を固める帯活動の試み― DDL を利用して」『関東甲信越英語教育学会第 48 回山梨研究大会』(2024 年 8 月 17 日, 於 都留文科大学).

Nishigaki, C., & Kakiba, A. (2023). What does data-driven learning (DDL) bring out in grammar learning? *Bulletin of the Faculty of Education, Chiba University*, *71*, 197–207.

Nishigaki, C., & Akasegawa, S. (2023). Development and revision of DDL tools for secondary school students: What we can do to nurture autonomous corpus users? *English Corpus Studies*, *30*, 131–149.

文部科学省（2017). 『中学校学習指導要領』.

文部科学省（2018). 『高等学校学習指導要領』.

Nation, P. (2024). *What should every EFL teacher know?* (2nd ed.). Compass Publishing.

白畑知彦（2015). 『英語指導における効果的な誤り訂正：第二言語習得研究の見地から』. 大修館書店.

Takahashi, S., & Fujiwara, Y. (2016). Effects of inductive learning based on data-driven learning at elementary schools in Japan. *JES Journal*, *16*(1), 84–99.

Yano, Y., Long, M. H., & Ross, S. (1994). The effects of simplified and elaborated texts on foreign language reading comprehension, *Language Learning*, *44*(2), 189–219.

hDDL https://h.ddl-study.org/

hDDL2 https://hddl.ddl-study.org/

（西垣知佳子　千葉大学）

（赤瀬川史朗　Lago NLP）

「ソフトウェア紹介」

Estimating the CEFR-J Level of English Reading Passages: Development and Accuracy of CVLA3

Satoru UCHIDA and Masashi NEGISHI

1. Introduction

Assessing the difficulty level of English texts is essential for effective, personalized education. Numerous applications such as Bax's (2012) Text Inspector and Mizumoto's (2022) New Word Level Checker have been developed, demonstrating the high demand for these tools. This paper reports on the CEFR-based Vocabulary Level Analyzer, Version 3 (CVLA3; https://cvla.langedu.jp/), designed to estimate the CEFR-J level of reading texts.

The previous version, CVLA2 (Uchida and Negishi, 2018), has been used in various studies (Azemoto & Uchida, 2022; Jodoi, 2023; Miura, 2021; Sato & Yamada, 2020). Feedback from these studies highlights the need for more stable assessment results, improved processing speeds, file-based processing, and the option for locally hosted versions. To address these needs, a new version was developed, with several enhancements. This study outlines the updates in CVLA3, followed by a report on the accuracy validation and comparative experiments with CVLA2.

2. Updates in CVLA3

2.1 Backend Update

In CVLA2, the TreeTagger is employed for backend processing, utilizing the treetaggerwrapper library in Python for part-of-speech (POS) analysis. Recently, spaCy, a native Python library, has been widely adopted, offering not only POS tagging, but also dependency parsing and named entity recognition with proven high performance and accuracy (cf. Altinok, 2021; Vasiliev, 2020). Considering potential future developments, such as local application deployment, CVLA3 has transitioned to an entirely Python-based backend. POS tagging and syntactic analysis leverage spaCy

3.7.2, with the en_core_web_sm dictionary. Additionally, textstat 0.7.4 was used to calculate the Automated Readability Index (ARI), which may result in differences from CVLA2's calculations. This update enables more accurate POS tagging and supports the integration of the new metrics introduced in subsequent sections.

2.2 Update to Training Data

The data used in CVLA2 were early CEFR-aligned materials published before 2013. Given the increased adoption and refined understanding of the CEFR in recent years, CVLA3 has shifted to the use of EFL textbooks published between 2014 and 2020 for statistical training. To ensure clear representation, textbooks spanning multiple levels, such as A1-A2 or A2-B1, were excluded. Instead, 539 texts specifically classified as A1, A2, B1, B2, or C1 were selected (a sufficient number of C2-level texts were unavailable). These were then randomly split, with 431 texts (80%) designated for training and 108 texts (20%) for evaluation testing. This update reflects a more current interpretation of CEFR levels and includes C1-level texts, an addition from the CVLA2 that only covers levels A1–B2.

2.3 Update to Metrics

In CVLA2, four metrics were used: AvrDiff (average difficulty of content words classified as A1 to B2 level), BperA (ratio of B-level to A-level content words), ARI (Automated Readability Index), and VperSent (average number of verbs per sentence) (for details, see Uchida and Negishi, 2018). The first two metrics represent the lexical complexity, whereas the latter two reflect the sentence and text complexities.

In CVLA3, AvrDiff was calculated by adding C1 (470 words) and C2 (381 words) words from the English Vocabulary Profile wordlist. Previously, C-level words were highlighted in red in the output, but they were not included in the calculation, which may have been confusing to users. Since the EVP C-level list is limited, its inclusion is not expected to have a significant impact on the calculation results (but improves interpretability). In addition, CVLA3 expanded the set of metrics to eight, adding CVV1, AvrFreqRank, POStypes, and LenNP, allowing for a more detailed analysis of English texts and potentially enhancing the accuracy of level estimation.

CVV1 is defined as "the number of verb tokens divided by the square root of twice the number of verbs" (Spring & Johnson, 2022) and has been validated as an effective measure for evaluating English writing. Essentially, this metric represents lexical diversity, particularly in verb use, an area not covered by CVLA2. Note that be-verbs were not included in this calculation. **AvrFreqRank** represents the average rank of words based on their frequency in the Corpus of Contemporary American English (COCA). Items ranked above 10,000 were uniformly calculated as 10,000 to prevent outliers. Additionally, the three most infrequent words were excluded from the overall calculations to compute the average. This approach minimizes the impact of low-frequency words, particularly when the passages are short. Unlike AvrDiff and BperA, which focus exclusively on content words, AvrFreqRank includes all the words, allowing for a comprehensive lexical-level analysis. Furthermore, it assigns unique values to each word based on rank rather than broad-level categories (AvrDiff calculates levels as 1 for A1, 2 for A2, and so on). Thus, CVV1 and AvrFreqRank offer more detailed assessments of lexical complexity.

POStypes is used to calculate the average number of distinct POS tags per sentence. More complex and longer sentences tend to include a wider range of tags, making a higher POS-type value indicative of greater grammatical complexity. Whereas VperSent focuses solely on verb counts, POStypes accounts for all parts of speech. **LenNP** represents the average length of the noun phrases calculated using spaCy POS tagging and dependency parsing. It measures the lengths of noun phrases that serve primarily as subjects or objects in a sentence. Longer noun phrases are presumed to increase the sentence difficulty, suggesting a higher level of complexity as LenNP increases. Together, POStypes and LenNP provide new perspectives on sentence complexity beyond those offered by the CVLA2 metrics.

2.4 Update to Evaluation Method

Considering the updates described above, CVLA3 assesses CEFR-J levels by utilizing new metrics in the updated corpus. Table 1 presents the average values for each metric across CEFR levels in the new textbook corpus, highlighting the linear trend in which each metric increased with higher levels. This linear relationship allows the construction of simple regression equations for each metric, providing a clear and interpretable framework. Therefore, users can easily identify the metrics that most strongly indicate higher or lower difficulty levels.

Based on the results in this table, we developed regression equations using the

Table 1. Average values of each metric by CEFR level

CEFR	AvrDiff	BperA	CVV1	AvrFreqRank	ARI	VperSent	POStypes	LenNP
A1	1.28	0.06	1.93	367.99	4.10	1.51	7.16	2.94
A2	1.44	0.12	2.95	445.92	6.22	2.05	8.14	3.36
B1	1.57	0.18	3.90	514.55	7.82	2.66	8.73	3.64
B2	1.74	0.26	4.67	613.05	9.19	2.95	9.04	3.99
C1	1.91	0.36	5.58	739.30	10.79	3.28	9.36	4.51

level assignments of A1 = 1, A2 = 2, B1 = 3, B2 = 4, and C1 = 5, following the approach used in CVLA2. To prevent outliers from skewing the results, an upper limit of 7 was applied to these equations.

CVV1_CEFR=min (CVV1×1.1059-1.208, 7)
BperA_CEFR=min (BperA×13.146+0.428, 7)
POStypes_CEFR=min (POStypes×1.768-12.006, 7)
ARI_CEFR=min (ARI×0.607-1.632, 7)
AvrDiff_CEFR=min (AvrDiff×6.417-7.184, 7)
AvrFreqRank_CEFR=min (AvrFreqRank×0.004-0.608, 7)
VperSent_CEFR=min (VperSent×2.203-2.486, 7)
LenNP_CEFR=min (LenNP×2.629-6.697, 7)

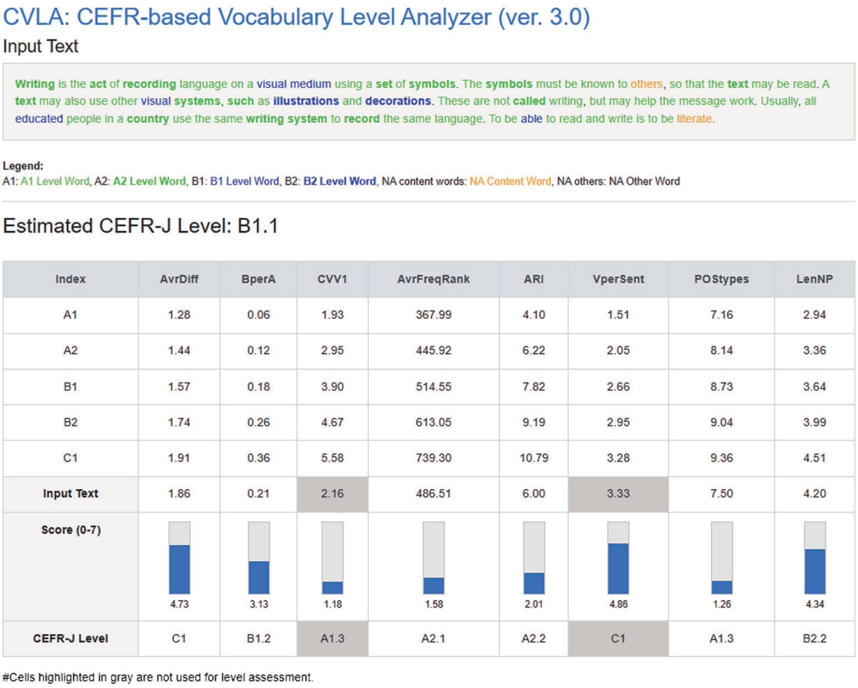
To ensure stability, the final value was calculated by excluding the minimum and maximum values from the regression results and averaging the six middle values. Notably, a raw metric value of zero (i.e. before CEFR conversion) is not necessarily excluded as the lowest value owing to the nature of the regression equation. For example, when BperA is zero, it yields a value of 0.428; therefore, if other values are lower, BperA will not be excluded.

The conversion to CEFR-J levels followed the method outlined by Uchida and Negishi (2018), as shown in Table 2. Figure 1 presents the sample analysis results, with gray-shaded metrics indicating those that were not used in the calculation.

For the sample text, the CEFR scores for each metric were AvrDiff = 4.73, BperA = 3.13, CVV1 = 1.18, AvrFreqRank = 1.58, ARI = 2.01, VperSent = 4.86, POStypes =

Table 2. Mapping to CEFR-J levels

Range	CEFR-J	Range	CEFR-J
$x < 0.5$	preA1	$2.5 \leq x < 3$	B1.1
$0.5 \leq x < 0.84$	A1.1	$3 \leq x < 3.5$	B1.2
$0.84 \leq x < 1.17$	A1.2	$3.5 \leq x < 4$	B2.1
$1.17 \leq x < 1.5$	A1.3	$4 \leq x < 4.5$	B2.2
$1.5 \leq x < 2$	A2.1	$4.5 \leq x < 5.5$	C1
$2 \leq x < 2.5$	A2.2	$x \geq 5.5$	C2



CVLA: CEFR-based Vocabulary Level Analyzer (ver. 3.0)

Input Text

Writing is the **act** of **recording** language on a visual medium using a **set** of **symbols**. The **symbols** must be known to others, so that the **text** may be read. A **text** may also use other visual **systems**, **such** as **illustrations** and **decorations**. These are not **called** writing, but may help the message work. Usually, all educated people in a **country** use the same **writing system** to **record** the same language. To be able to read and write is to be literate.

Legend:
A1: A1 Level Word, A2: **A2 Level Word**, B1: B1 Level Word, B2: **B2 Level Word**, NA content words: NA Content Word, NA others: NA Other Word

Estimated CEFR-J Level: B1.1

Index	AvrDiff	BperA	CVV1	AvrFreqRank	ARI	VperSent	POStypes	LenNP
A1	1.28	0.06	1.93	367.99	4.10	1.51	7.16	2.94
A2	1.44	0.12	2.95	445.92	6.22	2.05	8.14	3.36
B1	1.57	0.18	3.90	514.55	7.82	2.66	8.73	3.64
B2	1.74	0.26	4.67	613.05	9.19	2.95	9.04	3.99
C1	1.91	0.36	5.58	739.30	10.79	3.28	9.36	4.51
Input Text	1.86	0.21	2.16	486.51	6.00	3.33	7.50	4.20
Score (0-7)	4.73	3.13	1.18	1.58	2.01	4.86	1.26	4.34
CEFR-J Level	C1	B1.2	A1.3	A2.1	A2.2	C1	A1.3	B2.2

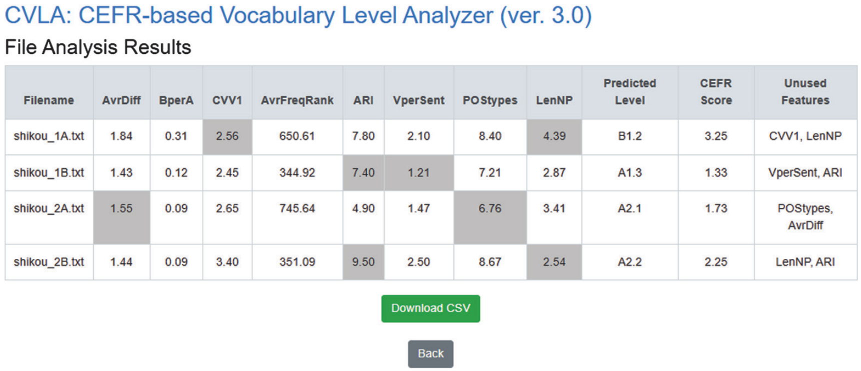
#Cells highlighted in gray are not used for level assessment.

Figure 1. Analysis results of sample text using CVLA3

1.26, and LenNP = 4.34. Excluding the minimum (CVV1 = 1.18) and maximum (VperSent = 4.86) values, the average of the six remaining values was 2.84. According to Table 2, this score corresponds to a CEFR of B1.1.

2.5 Addition of File Mode

To facilitate the processing of large volumes of files, CVLA3 includes a file mode that supports batch processing. Users can upload up to 30 text files with a maximum size of 10 KB per file. The results are output as a summary table, which can be downloaded in the CSV format, enabling efficient analysis of extensive datasets. Figure 2 shows a sample screen of the results generated in file mode.



CVLA: CEFR-based Vocabulary Level Analyzer (ver. 3.0)

File Analysis Results

Filename	AvrDiff	BperA	CVV1	AvrFreqRank	ARI	VperSent	POStypes	LenNP	Predicted Level	CEFR Score	Unused Features
shikou_1A.txt	1.84	0.31	2.56	650.61	7.80	2.10	8.40	4.39	B1.2	3.25	CVV1, LenNP
shikou_1B.txt	1.43	0.12	2.45	344.92	7.40	1.21	7.21	2.87	A1.3	1.33	VperSent, ARI
shikou_2A.txt	1.55	0.09	2.65	745.64	4.90	1.47	6.76	3.41	A2.1	1.73	POStypes, AvrDiff
shikou_2B.txt	1.44	0.09	3.40	351.09	9.50	2.50	8.67	2.54	A2.2	2.25	LenNP, ARI

Figure 2. Example of results in the file mode

3. Accuracy Validation

This section reports the accuracy of the CVLA3. Although CVLA3 was designed to estimate CEFR-J levels, no corpus with pre-assigned CEFR-J levels currently exists. Therefore, we conducted validation using texts labeled with standard CEFR levels, converting the levels as follows for consistency: preA1, A1.1, A1.2, and A1.3 were converted to A1; A2.1 and A2.2 to A2; B1.1 and B1.2 to B1; and B2.1, B2.2, to B2. In a previous study, the CVLA2 achieved an accuracy of approximately 53% on the CEFR scale (Uchida and Negishi, 2021).

The evaluation dataset used for the validation consisted of 108 English texts from an updated CEFR-aligned corpus. Table 3 shows the accuracy results of CVLA3, with rows representing the actual text levels and columns representing CVLA3's predicted levels. Of the 108 texts, CVLA3 correctly identified 71 cases (highlighted in dark blue), resulting in an accuracy of 65.74%. When accounting for adjacent levels (highlighted in light blue), the accuracy increased to 107 matches, indicating a high stability of 99.07%.

Table 3. Level estimation results of CVLA3 on the evaluation dataset

	A1	A2	B1	B2	C1	C2	total
A1	**16**	2					18
A2	2	**17**	3				22
B1		8	**14**	4			26
B2			5	**14**	2		21
C1			1	9	**10**	1	21
total	18	27	23	27	12	1	108

4. Comparison with CVLA2

Table 4 presents the validation results for CVLA2, using the same evaluation dataset. CVLA2 correctly identified 65 of 108 cases, yielding an accuracy rate of 60.19%. Although slightly lower than CVLA3's accuracy, this result reaffirms that CVLA2 still offers a practical level of accuracy for practical applications.

Table 5 presents a cross-tabulation of the results based on CEFR-J levels using the same dataset. The match rate at the CEFR level (six categories, highlighted in light blue) was 77 out of 108 (71.30 %). For CEFR-J levels (12 categories, highlighted in dark blue), the match rate was 55 of 108 (50.93 %). Although the judgment results may vary depending on the CVLA versions, the validation results and increased number of metrics suggest that CVLA3 is likely to provide higher accuracy and greater stability.

Table 4. Level estimation results of CVLA2 on the evaluation dataset

	A1	A2	B1	B2	C1	C2	total
A1	**16**	2					18
A2	6	**15**	1				22
B1		9	**12**	4	1		26
B2			4	**13**	4		21
C1			1	7	**9**	4	21
total	22	26	18	24	14	4	108

Table 5. Comparison of CEFR-J level estimation results between CVLA2 (row) and CVLA3 (column)

	preA1	A1.1	A1.2	A1.3	A2.1	A2.2	B1.1	B1.2	B2.1	B2.2	C1	C2	total
preA1	**4**	2	1										7
A1.1	2		1	1									4
A1.2				1	1								2
A1.3			1	**2**	6								9
A2.1				3	**7**	3	1						14
A2.2					2	**6**	4						12
B1.1						1	**5**	1					7
B1.2						1		**9**	1				11
B2.1								3	**9**	4			16
B2.2									3	**4**	1		8
C1										6	**8**		14
C2											3	**1**	4
total	6	2	3	7	16	11	10	13	13	14	12	1	108

5. Conclusion and Future Directions

CVLA3 has achieved substantial enhancements through updates to its backend, corpus foundation, metrics, evaluation methods, and the addition of a file mode, resulting in a faster and more stable web application. With an accuracy rate of 65.74%

in predicting CEFR levels (6-level classification) using the evaluation dataset, it serves as a valuable tool for assessing text difficulty.

The listening mode was not implemented in this revision because of challenges such as insufficient data and the need for audio-based metrics, such as Words Per Minute, for accurate assessment. However, incorporation of this feature should be considered in future studies. Additionally, we have released a beta version of the local application (currently available for Windows only), which allows users to analyze sensitive data offline. Further refinement may be required based on user feedback.

References

Altinok, D. (2021). *Mastering spaCy: An end-to-end practical guide to implementing NLP applications using the Python ecosystem*. Packt Publishing Ltd.

Azemoto, R., & Uchida, S. (2022). The importance of criterial features for CEFR-based textbooks: A case study using CVLA. *Studies in Languages and Cultures*, *48*, 35–47. https://doi.org/10.15017/4773109［畔元里沙子・内田諭. (2022).「CEFR レベル別英語教科書における基準特性の重要度 : CVLA の指標を用いて」『言語文化論究』48, 35–47.］

Bax, S. (2012). Text Inspector: Online text analysis tool. Available at: https://textinspector.com/.

Jodoi, K. (2023). The correlations between parliamentary debate participation, communication competence, communication apprehension, argumentativeness, and willingness to communicate in a Japanese context. *Argumentation*, *37*(1), 91–118. https://doi.org/10.1007/s10503-022-09591-5

Miura, A. (2021). Identifying the CEFR-J Levels of the Reading Texts Introduced in a Course for Current English 1 (Reading). *Journal of Multilingual Pedagogy and Practice*, *1*, 1–15. https://doi.org/10.14992/00020483

Mizumoto, A. (2022). An overview of New Word Level Checker. *Proceedings of the Methodology Study Group*, *12*, 1–24. https://doi.org/10.69194/methodologysig.12［水本篤. (2022).「New Word Level Checker の概要」『メソドロジー研究部会報告論集』12, 1–24.］

Sato, T., & Yamada, Y. (2020). Comparison of the text difficulty in new university entrance examinations. *Bulletin of the Chubu English Language Education Society*, *49*, 149–156. https://doi.org/10.20713/celes.49.0_149［佐藤選・山田裕也. (2020).「新大学入試におけるリーディング文章の難易度比較」『中部地区英語教育学会紀要』49, 146–156.］

Spring, R., & Johnson, M. (2022). The possibility of improving automated calculation of measures of lexical richness for EFL writing: A comparison of the LCA, NLTK and

SpaCy tools. *System*, *106*, 102770. https://doi.org/10.1016/j.system.2022.102770

Uchida, S., & Negishi, M. (2018). Assigning CEFR-J levels to English texts based on textual features. In Y. Tono and H. Isahara (eds.) *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference* (APCLC 2018), pp. 463–467.

Uchida, S., & Negishi, M. (2021). Estimating the CEFR levels of English reading materials: Evaluation of CVLA. *Journal of Corpus-Based Lexicology Studies*, *3*, 1–14.［内田諭・根岸雅史（2021）「英語読解教材の CEFR レベルの推定：CVLA の妥当性評価」*Journal of Corpus-based Lexicology Studies*, *3*, 1–14.］

Vasiliev, Y. (2020). *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press.

（Satoru Uchida, Kyushu University Email: uchida@flc.kyushu-u.ac.jp）
（Masashi Negishi, Tokyo University of Foreign Studies）

英語コーパス研究（第32号）

【 2025年5月31日発行 】

編集・発行　©2025　英語コーパス学会
〒464-8601 名古屋市千種区不老町
名古屋大学言語教育センター　小島ますみ研究室
E-mail: jaecs.hq@gmail.com
Website: https://jaecs.com/
郵便振替口座：009300-3-195373 （英語コーパス学会）
印　刷　所　株式会社ソウブン・ドットコム
〒116-0011　東京都荒川区西尾久 7 丁目 12 番 16 号

ISSN 1340-301 X

English Corpus Studies: Vol.32
2025

Research Articles

Research Note

Software Introduction

JAPAN ASSOCIATION FOR ENGLISH CORPUS STUDIES