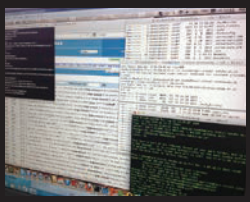
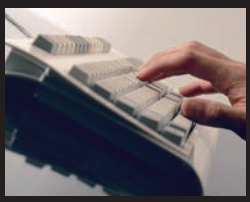
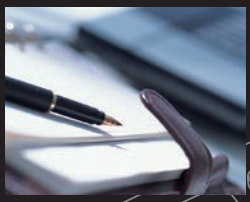
JAECS
Japan Association for English Corpus Studies
2022
英語コーパス学会

ISSN 1340-301 X

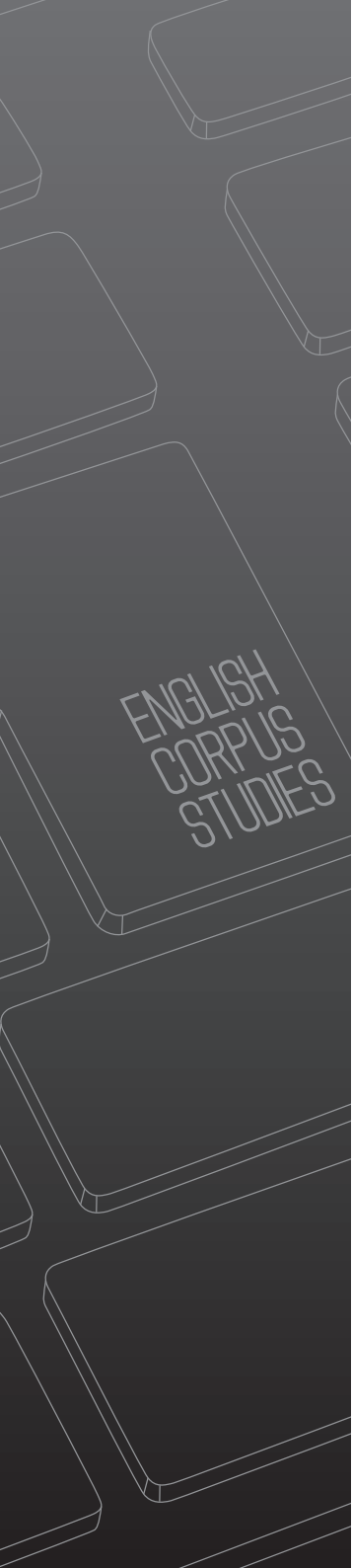






英語コーパス研究
第29号

English Corpus Studies 29



2022年度
英語コーパス学会役員

会　　長：田畑　智司

副 会 長：家入　葉子（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：阿部真理子，後藤　一章（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：家入　葉子，石井　康毅，石川　保茂，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，杉森　直樹，住吉　　誠，
　　田畑　智司，塚本　　聡，投野由紀夫，水野　和穂，
　　水本　　篤，山﨑　　聡，山本史歩子

幹　　事：阿部真理子，五百蔵高浩，石川　有香，和泉　絵美，
　　宇佐美裕子，木山　直毅，後藤　一章，佐竹　由帆，
　　長　加奈子，西垣知佳子，仁科　恭徳，藤原　康弘，
　　森下　裕三，渡辺　拓人

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＞今林　　修，大谷　直輝，木山　直毅，福元　広二，
　　三浦　愛香，水本　　篤，山﨑のぞみ

学会賞選考委員会
＜委 員 長＞塚本　　聡
＜委　　員＞石川　保茂，今林　　修，高橋　　薫，谷　　明信，
　　塚本　　聡，西垣知佳子

大会実行委員会
＜委 員 長＞藤原　康弘
＜委　　員＞鎌倉　義士，小島ますみ，James Rogers

ISSN 1340-301 X

英語コーパス研究

第 29 号

英語コーパス学会

2022

目　次

「論文」

Exploring L2 Spoken Developmental Measures: Which Linguistic Features Can Predict the Number of Words?

Yuichiro KOBAYASHI, Mariko ABE, and Yusuke KONDO

Abstract

One of the challenges for research in second language (L2) acquisition is finding reliable indices to objectively measure language development. To this end, researchers usually compare language learners of different proficiency levels through language proficiency tests. However, these proficiency levels can vary because each proficiency scale has different objectives and evaluation criteria. If the levels to be compared change, the developmental indices identified in the comparison change accordingly. Considering these issues, we seek to explore the effectiveness of criteria other than test scores and proficiency levels. Statistically, word tokens can be an alternative measure of spoken proficiency levels, as there is a high correlation between speaking proficiency and the number of words used in L2 speech. In addition, word tokens can be measured objectively and more consistently than proficiency levels. The number of words need not be converted from test scores, as it can be directly calculated from learners' spoken performance. Given these advantages, the present study investigates the mechanism of the increase in word tokens in L2 speaking. To do this, we counted the frequencies of Biber's (1988) 67 linguistic features in 832 L2 speech samples. Using these frequencies as predictor variables for random forest regression analysis, the study identified the features that contribute to an increase in the number of words. The results suggest that (a) causative adverbial subordinators, (b) independent clause coordination, (c) emphatics, (d) nouns, (e) prepositional phrases, and (f) present tense can best predict language development. These six key features can be robust indices of spoken language progress because they are frequently used in almost all speaking situations. The findings of the current study also offer valuable new insights into the methodology of L2 developmental studies.

『英語コーパス研究』第29号（2022），pp. 1–18

1. Introduction

Indices that effectively and objectively measure language development are of notable challenge in the field of second language acquisition (SLA) research. To tackle this challenge, researchers have compared various linguistic characteristics, such as lexical diversity and syntactic complexity, extracted from performances by learners with different proficiency levels (Crossley & McNamara, 2012; Díez-Bedmar & Pérez-Paredes, 2020; Kyle et al., 2021; Kyle & Crossley, 2018; Lu, 2011; Tracy-Ventura et al., 2021; Verspoor et al., 2021; Vyatkina, 2013). However, the estimated proficiency levels based on the scores of different language tests can vary because different language tests have different objectives and evaluation criteria. For example, in a particular test, the number of grammatical errors is a good predictor of the estimated proficiency levels; however, in another test, it can be a poor predictor of the proficiency levels. Considering this issue, the effectiveness of criteria other than test scores and proficiency levels must be explored.

Word tokens have shown promise as an alternative measure of spoken proficiency levels. Statistically, there is a high correlation between speaking proficiency and the number of words in second language (L2) speech (Kobayashi & Abe, 2016; Kobayashi et al., 2018). In the initial stages of language acquisition, an increase in running words in a limited amount of time can be one of the best indicators of language development. In the later stages, the number of words can reflect syntactic complexity in L2 speech. In other words, we can assume that word tokens are an objective and consistent measurement of L2 speaking ability. In this study, we investigate the strength of word tokens as a measuring tool with the aim of seeing how we can use it as a valuable index.

2. Background

2.1 L2 Developmental Measures

Since the 1970s, SLA researchers have sought the best “yardstick” to measure L2 development (Larsen-Freeman, 1978). Traditionally, they have focused on the T-unit (Hunt, 1970) and the average length of error-free T-units (Larsen-Freeman & Strom, 1977) as developmental indices for L2 writing. In line with these studies, Wolfe-

Quintero et al. (1998) suggested that T-unit length, error-free T-unit length, and clause length can be considered the best measures for fluency. Since then, a number of developmental studies have investigated the dimensions of complexity, accuracy, and fluency (CAF or CALF when lexis is seen as an independent domain), to assess the quality of L2 speech and writing (Housen et al., 2012). While T-unit and CAF measures have been widely used in SLA studies, the debate about their validity and universality continues (Ortega, 2003; Norris & Ortega, 2009).

L2 developmental studies have greatly benefited from learner corpus research (LCR). The availability of learner corpora enables language researchers to empirically track the language acquisition process. Additionally, the development of natural language processing technology has made it possible to analyze a broad range of linguistic features as well as several types of language errors that occur in corpora. For example, Garner and Crossley (2018) examined the growth of n-gram use in multiple indices (frequency, association strength, proportion) in the spoken performance of L2 speakers over a period of four months; they subsequently demonstrated that the frequency and proportion of bigrams were strongly related to the learners' proficiency levels. Kyle and Crossley (2018) compared traditional indices of syntactic complexity (e.g., mean length of T-units), fine-grained indices of clausal complexity, and fine-grained indices of phrasal complexity, and showed that fine-grained indices of phrasal complexity were better predictors of L2 writing quality than the other two indices. Díez-Bedmar and Pérez-Paredes (2020) analyzed noun phrase syntactic complexity in L2 writing and suggested that *nouns and modifiers* and *determiner + multiple premodification + head* can be used as indices of syntactic complexity. Meunier and Littré (2013) tracked learners' longitudinal progress in the acquisition of the English tense and aspect system and reported that tense and aspect errors decrease over time. Thewissen (2013) investigated more than 40 types of errors in essays written by learners with different proficiency levels and indicated that there is a difference in the error patterns between B1 and B2 levels of the Common European Framework of Reference for Languages (CEFR). Other learner corpus studies have explored various developmental indices, such as pragmalinguistic features (Miura, 2020) and metadiscourse markers (Kobayashi, 2017), from the perspectives of pragmatics and discourse analysis respectively. However, most studies on developmental indicators have focused on L2 writing, with fewer based on L2 speaking.

2.2 Biber's Linguistic Features

The linguistic features used by Biber (1988) aid in providing a comprehensive description of L2 speaking development. His selected set of linguistic features is broadly used in corpus-based studies to explore various types of linguistic variation (Conrad & Biber, 2001; Frignal, 2013; Sardinha & Pinto, 2014, 2019). This trend can be applied to learner corpus studies to help in identifying linguistic features that can predict the development of learners' speech (Abe, 2014), and automatically assess L2 spoken performance (Kobayashi & Abe, 2016). In this study, 67 linguistic features from Biber (1988) were used as variables to predict the increase of words in L2 spoken performance. As Table 1 shows, these features can be classified into 16 major grammatical categories: (a) tense and aspect markers, (b) place and time adverbials, (c) pronouns and pro-verbs, (d) questions, (e) nominal forms, (f) passives, (g) stative forms, (h) subordination, (i) prepositional phrases, adjectives, and adverbs, (j) lexical specificity, (k) lexical classes, (l) modals, (m) specialized verb classes, (n) reduced forms and dispreferred structures, (o) coordination, and (p) negation. Given the diversity of linguistic features to be considered, high-dimensional statistical methods that can handle a large number of variables and identify a smaller number of important variables among many features are needed.

Table 1. The 67 linguistic features from Biber (1988)

A. Tense and aspect markers
1. past tense (VBD), 2. perfect aspect (PEAS), 3. present tense (VPRT)
B. Place and time adverbials
4. place adverbials (PLACE), 5. time adverbials (TIME)
C. Pronouns and pro-verbs
6. first person pronouns (FPP1), 7. second person pronouns (SPP2), 8. third person personal pronouns (excluding *it*) (TPP3), 9. pronoun *it* (PIT), 10. demonstrative pronouns (DEMP), 11. indefinite pronouns (INPR), 12. pro-verb *do* (PROD)
D. Questions
13. direct WH-questions (WHQU)
E. Nominal forms
14. nominalizations (ending in *-tion, -ment, -ness, -ity*) (NOMZ), 15. gerunds (GER), 16. total other nouns (NN)
F. Passives
17. agentless passives (PASS), 18. *by*-passives (BYPA)
G. Stative forms
19. *be* as main verb (BEMA), 20. existential *there* (EX)
H. Subordination features

21. *that* verb complements (THVC), 22. *that* adjective complements (THAC), 23. WH clauses (WHCL), 24. infinitives (*to*-clause) (TO), 25. present participial clauses (PRESP), 26. past participial clauses (PASTP), 27. past participial WHIZ deletion relatives (WZPAST), 28. present participial WHIZ deletion relatives (WZPRES), 29. *that* relative clauses on subject position (TSUB), 30. *that* relative clauses on object position (TOBJ), 31. WH relatives on subject position (WHSUB), 32. WH relatives on object position (WHOBJ), 33. pied-piping relative clauses (PIRE), 34. sentence relatives (SERE), 35. causative adverbial subordinators (*because*) (CAUS), 36. concessive adverbial subordinators (*although, though*) (CONC), 37. conditional adverbial subordinators (*if, unless*) (COND), 38. other adverbial subordinators (OSUB)

I. Prepositional phrases, adjectives, and adverbs

39. total prepositional phrases (PIN), 40. attributive adjectives (JJ), 41. predicative adjectives (PRED), 42. total adverbs (RB)

J. Lexical specificity

43. type/token ratio (TTR), 44. mean word length (AWL)

K. Lexical classes

45. conjuncts (CONJ), 46. downtoners (DWNT), 47. hedges (HDG), 48. amplifiers (AMP), 49. emphatics (EMPH), 50. discourse particles (DPAR), 51. demonstratives (DEMO)

L. Modals

52. possibility modals (POMD), 53. necessity modals (NEMD), 54. predictive modals (PRMD)

M. Specialized verb classes

55. public verbs (PUBV), 56. private verbs (PRIV), 57. suasive verbs (SUAV), 58. *seem* and *appear* (SMP)

N. Reduced forms and dispreferred structures

59. contractions (CONT), 60. subordinator *that* deletion (THATD), 61. stranded prepositions (STPR), 62. split infinitives (SPIN), 63. split auxiliaries (SPAU)

O. Coordination

64. phrasal coordination (PHC), 65. independent clause coordination (ANDC)

P. Negation

66. syntactic negation (SYNE), 67. analytic negation (XX0)

Note. The abbreviations given in parentheses are the tags used in the Multidimensional Analysis Tagger (Nini, 2019).

2.3 Multifactorial Regression Analysis

A new methodological trend in LCR is multifactorial regression analysis (Gries, 2015; Gries & Deshors, 2014, 2021; Gries & Wulff, 2013; Wulff & Gries, 2015, 2019, 2021). In this statistical method, multiple variables (e.g., linguistic features, language errors) can be used to determine the behavior of a response (e.g., proficiency levels, word tokens). Moreover, it can evaluate the strength of association between the predictors and response in the context of statistical significance tests (e.g., *t*-test, Wald test). Thus, it allows us to simultaneously assess the multiple factors involved in language development without repeating mono-factorial tests. Multifactorial regression

analysis can be broadly divided into linear and nonlinear models, depending on the types of fitting methods. Linear models presuppose a linear relationship between predictor and response variables, while nonlinear models formulate various nonlinear relationships between predictor and response variables in cases where a linear relationship cannot be assumed. While linear models have one basic form (i.e., *response* = *constant* + *parameter* * *predictor* + ... + *parameter* * *predictor*), nonlinear models can take many different forms. In the SLA context, the language development process is not linear (Murakami, 2016). Specifically, in U-shaped development, the learners' accuracy is high in the beginning, but it drops temporarily before increasing again. In addition, in power-law development, the decrement in error becomes gradually smaller as the learner's proficiency increases. With the awareness of the nonlinearity in SLA, Murakami (2016) applied generalized additive mixed models to investigate the nonlinear patterns of the L2 accuracy development in English grammatical morphemes. Verspoor et al. (2021) also utilized generalized additive models to examine the nonlinear development in the mean length of T-units and the Guiraud index.

Random forest (Breiman, 2001) is one of the most powerful multifactorial techniques for analyzing such nonlinear developmental patterns. The method is an ensemble learning technique that operates by constructing a large collection of regression trees. The regression tree model is a nonlinear regression technique that visualizes a sequence of data classification in the form of a flowchart-like diagram (Breiman et al., 1984). In the random forest model, the ensemble of regression trees (the forest) is generated using the ensemble learning technique, to yield better predictive performance than can possibly be obtained from any of the constituent tree models. The bagging ensemble learning algorithm (Breiman, 1994) is widely used to synthesize multiple tree models. It generates a number of datasets using a bootstrap sampling technique, and then constructs multiple regression models based on each bootstrap sample. Following these steps, the random forest model calculates the average of the predictions of every single regression tree to make a final prediction. By combining regression tree and bagging ensemble learning techniques, the random forest model generally achieves higher levels of predictions than other machine learning techniques (Chen et al., 2020). This model can also handle thousands of predictor variables in a statistically efficient manner (e.g., bootstrap sampling, feature

sampling), as it is more robust to multicollinearity than linear regression and several other regression models. Moreover, this model can compute variable importance scores to measure the impact of each predictor variable on the alternation, given all other predictors. Because of these advantages, random forest is regarded as a useful tool for the identification of L2 developmental indices.

The use of random forest models has been increasing in the field of corpus linguistics. For instance, Tono (2013) applied this technique to investigate several types of language errors that occur in L2 writing and found that the omission errors of *have* and *want* are the two most important predictors of English proficiency levels. Additionally, Kobayashi and Abe (2016) predicted the quality of L2 speech using random forest and showed that word tokens and types are the best predictors of speaking proficiency. In addition to these learner corpus studies, random forest has been utilized for studies in language usage, such as verb-object-particle vs. verb-particle-object alternation (Deshors, 2019), and the choice between the progressive and simple aspects (Hundt et al., 2020).

3. Purpose of the Study

As mentioned above, word tokens can be an alternative measure of L2 speaking proficiency from a statistical perspective. Therefore, adequate predictors of word tokens in learners' spoken performance can help SLA researchers in understanding proficiency. Against this background, the present study aimed to investigate the mechanism of the increase in word tokens in L2 speaking. The research questions (RQ) that drive this article are as follows:

RQ 1: How highly correlated is the number of words with L2 speaking proficiency?

RQ 2: Which linguistic features can contribute to an increase in the number of words in L2 speech?

By pursuing RQ 1, this study validates the effectiveness of the number of words as developmental measure for L2 speaking. In addition, the answer to RQ 2 can contribute to L2 speaking assessment including automated speech scoring.

4. Methods

4.1 Corpus

The spoken data utilized in this study were extracted from the Longitudinal Corpus of L2 Spoken English (LOCSE; Abe & Kondo, 2019). LOCSE was designed to describe L2 developmental patterns, not only at the group level, but also at the individual level. The speech samples were collected from upper-secondary school students. They were public senior high school students aged 15 years at the beginning of data collection. The students spoke Japanese as their mother tongue and had no long-term experience in English-speaking countries. Additionally, they were studying the target language under a similar learning setting and had limited opportunities to speak the target language inside and outside the classroom.

The students were asked to take a monologue speaking test, the Telephone Standard Speaking Test (TSST), which consists of multiple tasks (e.g., description, comparison, reasoning). Their utterances were compiled to create the corpus data. The automated telephone-based English-speaking test consists of ten recorded questions, and test-takers were required to respond to each question in 45 seconds without any planning time or use of reference material. Three certified raters gave a holistic score to each speech sample, based on various criteria such as function-based ability, sentence structure, accuracy, and content. The test scores were divided into nine levels, ranging from level 1 (novice) to level 9 (advanced).

The speech samples collected in the test were transcribed by four trained transcribers using automated speech recognition technology (IBM Watson Speech-to-Text). For the transcription, the XML format was chosen for the interchangeability of the resource, and the annotation schema of Izumi et al. (2004) was used for comparison with other learner corpora (e.g., the NICT-JLE Corpus, Konan-JIEM Learner Corpus, KIT Speaking Test Corpus).

This study analyzed speech samples from 104 students (47 boys and 57 girls) who had taken all eight speaking tests, making a total of 832 samples. However, this study did not make use of longitudinal information of this learner corpus. Table 2 summarizes the numbers and percentages of speech samples and words for each speaking proficiency level. As the table indicates, all learners were classified into TSST levels 2–7, which correspond to the CEFR levels A1–B1. As mentioned, this study used

the number of words as a criterion for assessing language development instead of proficiency levels (for an approach that uses proficiency as a criterion, refer to Kobayashi et al., 2018; for the longitudinal analysis of the LOCSE data, refer to Abe & Kondo, 2019).

Table 2. The numbers of speech samples and words in the LOCSE

TSST level	Number of speech samples		Number of words	
2	8	(0.96%)	762	(0.21%)
3	204	(24.52%)	63,313	(17.07%)
4	468	(56.25%)	207,654	(55.99%)
5	122	(14.66%)	75,836	(20.45%)
6	27	(3.25%)	20,835	(5.62%)
7	3	(0.36%)	2,485	(0.67%)
Total	832	(100.00%)	370,885	(100.00%)

4.2 Text Preprocessing

Before analyzing the transcribed speech samples, text preprocessing was conducted. Specifically, (a) fillers (e.g., *ah, eh, umm*), (b) Japanese words excluding proper nouns (e.g., *desu, kore, nandaro*), (c) words that the transcribers could not easily identify, (d) non-verbal phenomena (e.g., cough, laughter, sigh), (e) repetitions (e.g., *he he*), and (f) self-corrections of two words or less (e.g., *I I don't like cats but I like I like dogs*) were deleted. By removing these utterances, we can count learners' pruned tokens without dysfluency markers. Furthermore, this preprocessing can increase the accuracy of natural language processing, including part-of-speech tagging and syntactic parsing.

4.3 Data Analysis

This study counted the frequencies of Biber's (1988) linguistic features using the Multidimensional Analysis Tagger (Nini, 2019) and used the frequencies for correlation analysis and random forest regression analysis. All statistical analyses in this study were conducted using R, a free software environment for statistical computing and graphics (R Core Team, 2020). The *randomForest* package (Liaw & Wiener, 2002) was used to perform the analysis. For other R techniques, including correlation analysis and data visualization, this study mainly referred to Baayen (2008) and Levshina (2015).

5. Results

5.1 Correlation Analysis

The current study begins by investigating the correlation between learners' TSST levels and word tokens using Spearman's rank correlation coefficient. As a result, speaking proficiency was found to be highly correlated with the number of words in L2 spoken performance (ρ = 0.73). This means that word tokens can function as an alternative measure for TSST levels.

As a next step, we checked the correlations among Biber's linguistic features using Pearson's product-moment correlation coefficient. Table 3 lists the 20 pairs with the highest correlations. As the table shows, the highest correlation pair among features is contraction (CONT) and analytic negation (XX0) (r = 0.76), followed by *be* as main verb (BEMA) and predicative adjectives (PRED) (r = 0.69), and subordinator *that* deletion (THATD) and private verbs (PRIV) (r = 0.64). According to the correlation coefficients, the mean word length (AWL) in L2 speech increased with the number of nouns (NN) (r = 0.42) and fell with the repetition of first-person pronouns (FPP1) (r = −0.35). The type/token ratio (TTR) also decreased through the frequent use of first-person pronouns (r = −0.36).

Table 3. The 20 highest correlation pairs of linguistic features

Rank	Variable 1	Variable 2	r	Rank	Variable 1	Variable 2	r
1	CONT	XX0	0.76	11	BEMA	PIT	0.34
2	BEMA	PRED	0.69	12	PRED	PIT	0.32
3	THATD	PRIV	0.64	13	VPRT	AMP	0.31
4	BEMA	VPRT	0.45	14	PRED	AMP	0.31
5	NN	AWL	0.42	15	PHC	NN	0.31
6	VPRT	VBD	−0.36	16	EMPH	AMP	−0.31
7	FPP1	TTR	−0.36	17	FPP1	EMPH	−0.31
8	VPRT	PIN	−0.36	18	PIN	EMPH	0.31
9	FPP1	AWL	−0.35	19	JJ	AWL	0.29
10	VPRT	PRED	0.35	20	RB	NN	−0.29

5.2 Random Forest Regression Analysis

Given the high correlation of several of the pairs shown in Table 3, this study performed a random forest regression analysis that is relatively robust to

multicollinearity in the prediction. The random forest model used Biber's linguistic features as predictor variables and word tokens as the response variable. While running the statistical algorithm, the hyperparameters of the model (e.g., the number of trees and predictor variables randomly sampled as candidates for each tree) were tuned through the *tuneRF* function of the *randomForest* package. As a result of the tuning, the model generated 500 trees using 22 variables each and explained 58.73% of the total variance of the data.

The random forest model also estimated the importance of predictor variables using the increased node impurity index (IncNodePurity). Figure 1 shows the top 30 important linguistic features in the prediction of learners' word tokens. Variables that could predict the number of words in L2 speech were, in order of strength, frequency of causative adverbial subordinators (CAUS), independent clause coordination (ANDC), emphatics (EMPH), nouns (NN), prepositional phrases (PIN), and present tense (VPRT).

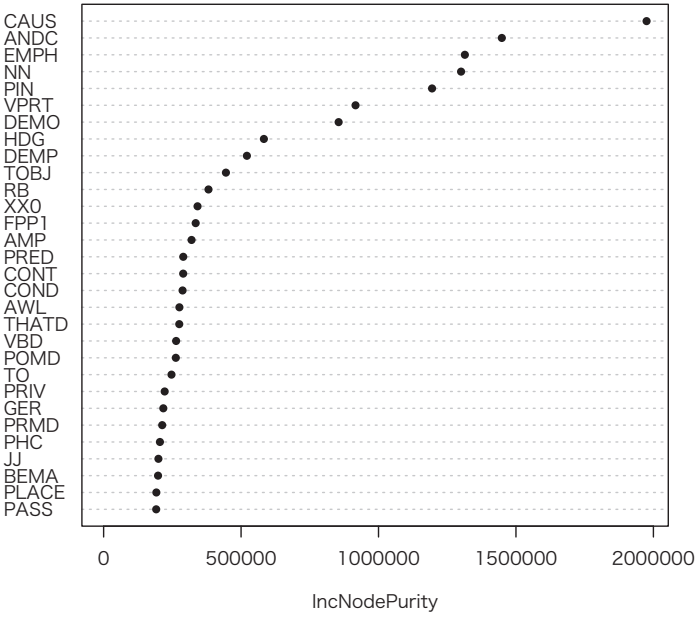


Figure 1. Variable importance plot of the top 30 linguistic features

Although there is no theoretical threshold that can be used to discriminate between important and unimportant variables, this study focuses on the top six linguistic features for detailed analysis. Figure 2 presents partial dependence plots that show how these six features affect the prediction of word tokens by marginalizing (averaging) out the effects of other features. By checking partial dependence plots, in addition to the variable importance plot, we can investigate the predictor variables while controlling for the effects of other variables (Hastie et al., 2009). The horizontal axes in the plots indicate the relative frequency of a particular linguistic feature (per 100 words), while the vertical axes indicate the number of tokens. As these plots illustrate, CAUS, NN, and VPRT are negatively related to word tokens, while EMPH and PIN are positively related. Additionally, the relative frequency of ANDC increases rapidly to around 0.1 and then decreases rapidly before it stabilizes, and it can discriminate learners in a specific range of word tokens. Interpreting the pattern in ANDC is more difficult than the patterns in the other items, but this is not because of a problem with our data. When predicting some natural phenomena, there are not many predictor variables that have values directly or inversely proportional to the values of the response variable. In the case of L2 assessment, there are some predictor variables that discriminate between learners who are above a certain level and those who are

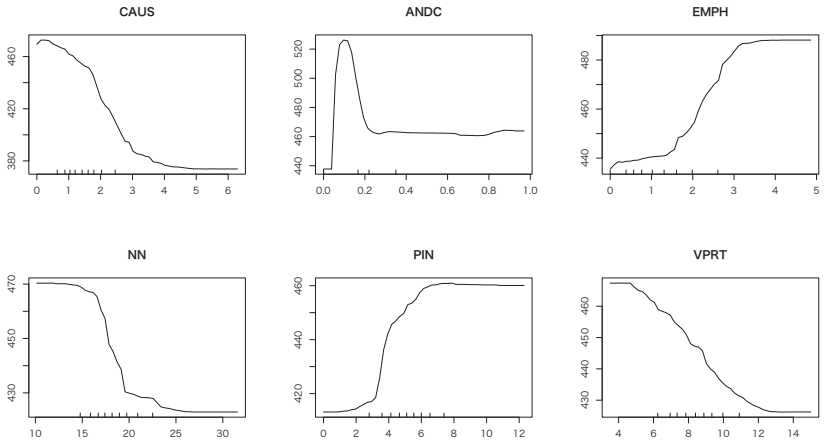


Figure 2. Partial dependence plots of the top six linguistic features

below a certain level. There are also some predictor variables that discriminate only between certain levels, such as the ANDC in this study. In other words, the random forest model provides highly accurate predictions by integrating the information held by these variables.

6. Discussion

Following the calculation of variable importance and partial dependence scores, this section explores the six important features that can predict the number of words in learners' utterances. The validity of these developmental indices will be further supported by checking concordance lines. First, the decrease in CAUS is attributed to the diversification of conjunctions that learners can use. As proficiency increases, learners can progressively construct speech without relying on the subordinating conjunction *because*. In other words, they move from the stage of "giving a reason" (e.g., *I like rainy day because rainy day is cool*) to the stage of "stating a result" (e.g., *Rainy day is cool, so I like rainy day*). Second, novice learners use ANDC with high frequency (e.g., *My mother is very careful woman, and she can find a lot of my mistakes, and she always advise me to improve my something, so I'm very I owe to her to improve my power of academic skills, and I'm very grateful for her*). After this stage, they will be able to use concessive adverbial subordinators (CONC), conditional adverbial subordinators (COND), and other adverbial subordinators (OSUB). Third, the increase in EMPH (e.g., *really, just, most, more*) allows advanced learners to express the degree of certainty in propositions more clearly. This rhetorical device can be a developmental index for both the dialogue speaking test (the Standard Speaking Test; Kobayashi & Abe, 2016) and the monologue test used in this study (the TSST). Fourth, the high frequency of NN is a prominent feature among novice learners (e.g., *I study ... five subject ... English ... Japanese, Math, and ... Science, and ... also ...*). They heavily depend on nouns in the initial stage of learning, but gradually become able to employ a variety of word types (Tono, 2000). Fifth, the increase in PIN results from the development of noun phrase structure. Additionally, prepositions become more frequent owing to the acquisition of group prepositions (e.g., *a lot of, because of*). Lastly, the decrease in VPRT use is a consequence of the increase of other tense use (e.g., *enjoyed, experienced, happened, tried*). As Table 3 shows, the frequency of the present tense is

negatively correlated with that of the past tense ($r = -0.36$).

7. Conclusion

This study aimed to explore the mechanism underlying the increase in number of words in L2 speaking. The results show that word tokens can function as an L2 developmental measure that highly correlates with speaking proficiency ($\rho = 0.73$). The results also suggest (a) causative adverbial subordinators, (b) independent clause coordination, (c) emphatics, (d) nouns, (e) prepositional phrases, and (f) present tense from Biber's linguistic features best predict the language development. These six key features can be robust measures of L2 spoken development, as they are frequently used in almost all speaking contexts. In addition, this study scrutinized the effects of these features on the increase in word tokens, by checking partial dependence plots. However, this study has some limitations. First, the frequencies of linguistic features may be affected by the tasks and topics of the TSST. Thus, we should investigate the effects of tasks and topics on learners' performance using multilevel analysis in the future. Second, the target learners were limited to novice and intermediate Japanese learners of English. It would be desirable to investigate a wider range of L1 backgrounds and proficiency levels to gain a broader understanding of the increase in word tokens in L2 speech. Third, other linguistic features can be useful for modeling the development of L2 spoken English. In particular, lexical and grammatical errors highlight language development from different angles than Biber's framework (Abe, 2007). Finally, because random forest is based on ensemble learning, a full interpretation of the results is difficult. One possible solution to this problem is to use global surrogate models that are trained to approximate the predictions of random forest models (Gries, 2020). Despite these limitations, the findings of the current study offer valuable new insights into the mechanism of the number of words in learners' speech as well as enhancing the methodology of L2 developmental studies.

Acknowledgements

This research has been partly funded by Grants-in-Aid for Scientific Research Grant Numbers 18K00849, 20K00813, and 21K00660, which we gratefully acknowledge. We are also very grateful to all the participants and researchers for their

help in the compilation of the LOCSE.

References

Abe, M. (2007) "A Corpus-based Investigation of Errors across Proficiency Levels in L2 Spoken Production." *JACET Journal 44*: 1–14.

Abe, M. (2014) "Frequency Change Patterns across Proficiency Levels in Japanese EFL Learner Speech." *Journal of Applied Language Studies 8*, 3: 85–96.

Abe, M., and Y. Kondo (2019) "Constructing a Longitudinal Learner Corpus to Track L2 Spoken English." *Journal of Modern Languages 29*: 23–44.

Baayen, R. H. (2008) *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R.* Cambridge: Cambridge University Press.

Biber, D. (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Breiman, L. (1994) "Bagging Predictors." *Machine Learning 24*, 2: 123–140.

Breiman, L. (2001) "Random Forests." *Machine Learning 45*, 1: 5–32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984) *Classification and Regression Trees*. Boca Raton: Chapman and Hall.

Chen, R.-C., C. Dewi, S.-W. Huang, and R. E. Caraka (2020) "Selecting Critical Features for Data Classification Based on Machine Learning Methods." *Journal of Big Data 7*, 52: 1–26.

Conrad, S., and D. Biber (Eds.) (2001) *Variation in English: Multi-dimensional Studies*. London: Longman.

Crossley, S. A., and D. S. McNamara (2012) "Predicting Second Language Writing Proficiency: The Roles of Cohesion and Linguistic Sophistication." *Journal of Research in Reading 35*, 2: 115–135.

Deshors, S. C. (2019) "English as a Lingua Franca: A Random Forests Approach to Particle Placement in Multi-speaker Interactions." *International Journal of Applied Linguistics 30*, 2: 214–231.

Díez-Bedmar, M. B., and P. Pérez-Paredes (2020) "Noun Phrase Complexity in Young Spanish EFL Learners' Writing: Complementing Syntactic Complexity Indices with Corpus-driven Analyses." *International Journal of Corpus Linguistics 25*, 1: 4–35.

Frignal, E. (2013) "Twenty-five Years of Biber's Multi-dimensional Analysis: Introduction to the Special Issue and an Interview with Douglas Biber." *Corpora 8*, 2: 137–152.

Garner, J., and S. Crossley (2018) "A Latent Curve Model Approach to Studying L2 N-gram Development." *The Modern Language Journal 102*, 3: 494–511.

Gries, S. Th. (2015) "Statistics for Learner Corpus Research." In Granger, S., G. Gilquin and F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, pp. 159–182.

Gries, S. Th. (2020) "On Classification Trees and Random Forests in Corpus Linguistics:

Some Words of Caution and Suggestions for Improvement." *Corpus Linguistics and Linguistic Theory 16*, 3: 617–647.

Gries, S. Th., and S. C. Deshors (2014) "Using Regressions to Explore Deviations between Corpus Data and a Standard/target: Two Suggestions." *Corpora 9*, 1: 109–136.

Gries, S. Th., and S. C. Deshors (2021) "Statistical Analyses of Learner Corpus Data." In Tracy-Ventura, N., and M. Paquot (Eds.), *The Routledge Handbook of Second Language Acquisition and Corpora*. New York: Routledge, pp. 119–132.

Gries, S. Th., and S. Wulff (2013) "The Genitive Alternation in Chinese and German ESL Learners: Towards a Multifactorial Notion of Context in Learner Corpus Research." *International Journal of Corpus Linguistics 18*, 3: 327–356.

Hastie, T., R. Tibshirani, and J. Friedman (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Edition. New York: Springer.

Housen, A., F. Kuiken, and I. Vedder (2012) *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Amsterdam: John Benjamins.

Hundt, M., P. Rautionaho, and C. Strobl (2020) "Progressive or Simple? A Corpus-based Study of Aspect in World Englishes." *Corpora 15*, 1: 77–106.

Hunt, K. W. (1970) "Do Sentences in the Second Language Grow Like Those in the First?" *TESOL Quarterly 4*, 3: 195–202.

Izumi, E., K. Uchimoto, and H. Isahara (2004) *A Speaking Corpus of 1,200 Japanese Learners of English.* Tokyo: ALC Press.

Kobayashi, Y. (2017) "Developmental Patterns of Metadiscourse in Second Language Writing." *Journal of Pan-Pacific Association of Applied Linguistics 21*, 2: 41–54.

Kobayashi, Y., and M. Abe (2016) "Automated Scoring of L2 Spoken English with Random Forests." *Journal of Pan-Pacific Association of Applied Linguistics 20*, 1: 55–73.

Kobayashi, Y., Y. Kondo, and M. Abe (2018) "Predicting EFL Learners' Oral Proficiency Levels in Monologue Tasks." In Tono, Y., and H. Isahara (Eds.), *Proceedings of the 4th Asia Pacific Corpus Linguistic Conference*, pp. 231–236.

Kyle, K., and S. A. Crossley (2018) "Measuring Syntactic Complexity in L2 Writing Using Fine-grained Clausal and Phrasal Indices." *The Modern Language Journal 102*, 2: 333–349.

Kyle, K., S. Crossley, and M. Verspoor (2021) "Measuring Longitudinal Writing Development Using Indices of Syntactic Complexity and Sophistication." *Studies in Second Language Acquisition 43*, 4: 781–812.

Larsen-Freeman, D. (1978) "An ESL Index of Development." *TESOL Quarterly 12*, 4: 439–448.

Larsen-Freeman, D., and V. Strom (1977) "The Construction of a Second Language Acquisition Index of Development." *Language Learning 27*, 1: 123–134.

Levshina, N. (2015) *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins.

Liaw, A., and M. Wiener (2002) "Classification and Regression by randomForest." *R News 2*, 3: 18–22.

Lu, X. (2011) "A Corpus-based Evaluation of Syntactic Complexity Measures as Indices of College-level ESL Writers' Language Development." *TESOL Quarterly 45*, 1: 36–62.

Meunier, F., and D. Littré (2013) "Tracking Learners' Progress: Adopting a Dual 'Corpus cum Experimental Data' Approach." *The Modern Language Journal 97*, S1: 61–76.

Miura, A. (2020) "Criterial Pragmalinguistic Features of Requestive Speech Acts Produced by Japanese Learners of English." *Learner Corpus Studies in Asia and the World 4*: 1–23.

Murakami, A. (2016) "Modeling Systematicity and Individuality in Nonlinear Second Language Development: The Case of English Grammatical Morphemes." *Language Learning 66*, 4: 834–871.

Nini, A. (2019) "The Multi-Dimensional Analysis Tagger." In Sardinha, T. B., and M. V. Pinto (Eds.), *Multi-dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic, pp. 67–94.

Norris, J. M., and L. Ortega (2009) "Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity." *Applied Linguistics 30*, 4: 555–578.

Ortega, L. (2003) "Syntactic Complexity Measures and Their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing." *Applied Linguistics 24*, 4: 492–518.

R Core Team (2020) R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. http://www.r-project.org/

Sardinha, T. B., and M. V. Pinto (Eds.) (2014) *Multi-dimensional Analysis, 25 Years on: A Tribute to Douglas Biber*. Amsterdam: John Benjamins.

Sardinha, T. B., and M. V. Pinto (Eds.) (2019) *Multi-dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic.

Thewissen, J. (2013) "Capturing L2 Accuracy Developmental Patterns: Insights from an Error-tagged EFL Learner Corpus." *The Modern Language Journal 97*, S1: 77–101.

Tono, Y. (2000) "A Corpus-based Analysis of Interlanguage Development: Analysing Part-of-speech Tag Sequences of EFL Learner Corpora." In Lewandowska-Tomaszczyk, B., and J. P. Melia (Eds.), *PALC'99: Practical Applications in Language Corpora*. Frankfurt am Main: Peter Lang, pp. 323–340.

Tono, Y. (2013) "Criterial Feature Extraction Using Parallel Learner Corpora and Machine Learning." In Díaz-Negrillo, A., N. Ballier, and P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: John Benjamins, pp. 169–204.

Tracy-Ventura, N., A. Huensch, and R. Mitchell (2021) "Understanding the Long-term Evolution of L2 Lexical Diversity: The Contribution of a Longitudinal Learner Corpus." In Bruyn, B. L., and M. Paquot (Eds.), *Learner Corpus Research Meets Second Language Acquisition*. Cambridge: Cambridge University Press, pp. 148–171.

Verspoor, M., W. Lowie, and M. Wieling (2021) "L2 Developmental Measures from a Dynamic Perspective." In Bruyn, B. L., and M. Paquot (Eds.), *Learner Corpus Research*

Meets Second Language Acquisition. Cambridge: Cambridge University Press, pp. 172–190.

Vyatkina, N. (2013) "Specific Syntactic Complexity: Developmental Profiling of Individuals Based on an Annotated Learner Corpus." *The Modern Language Journal 97*, S1: 11–30.

Wolfe-Quintero, K., S. Inagaki, and H.-Y. Kim (1998) *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. Honolulu: University of Hawaii Press.

Wulff, S., and S. Th. Gries (2015) "Prenominal Adjective Order Preferences in Chinese and German L2 English: A Multifactorial Corpus Study." *Linguistic Approaches to Bilingualism 5*, 1: 122–150.

Wulff, S., and S. Th. Gries (2019) "Particle Placement in Learner Language." *Language Learning 69*, 4: 873–910.

Wulff, S., and S. Th. Gries (2021) "Exploring Individual Variation in Learner Corpus Research: Methodological Suggestions." In Bruyn, B. L., and M. Paquot (Eds.), *Learner Corpus Research Meets Second Language Acquisition*. Cambridge: Cambridge University Press, pp. 191–213.

（小林雄一郎　日本大学　Email: kobayashi.yuuichirou@nihon-u.ac.jp）
（阿部真理子　中央大学　Email: abe.127@g.chuo-u.ac.jp）
（近藤　悠介　早稲田大学　Email: yusukekondo@waseda.jp）

「論文」

A Learner Corpus-Based Study of L1 Effects on L2 English Auxiliary Verb Use —The Case of *Will*—

Laurence NEWBERY-PAYTON

Abstract

This study conducts Contrastive Interlanguage Analysis of written essays by L1 Chinese and L1 Japanese learners of English contained in the International Corpus Network of Asian Learners of English (ICNALE). The two groups of learners are compared with each other and with native speakers regarding their use of the modal auxiliary *will*. Consideration of relevant characteristics of Chinese and Japanese suggests that L1 Chinese learners will overuse *will* due to functional similarities with the Chinese modal auxiliary *huì*, whereas this trend is not predicted to occur among L1 Japanese learners. Analysis of the corpus data reveals that Chinese L1 learners do overuse *will* at lower proficiency levels, providing evidence for crosslinguistic influence. In contrast, Japanese L1 learners, who lack functional equivalents to *will* in their native language, exhibit underuse as well as omission in obligatory contexts. The study therefore confirms the hypothesis that at lower proficiency levels, the presence or absence in L1 of partial functional equivalents to a target form can affect the latter's frequency of use in L2. However, these trends are restricted to one of two essay tasks, suggesting task-related factors.

1. Introduction

It has long been recognized by researchers that acquisition of a foreign language can be influenced by learners' native languages as well as any other languages previously acquired (Jarvis & Pavlenko, 2008; Luk & Shirai, 2009). This paper uses corpus data to conduct Contrastive Interlanguage Analysis of writing by Chinese L1 and Japanese L1 learners of English. In particular, it attempts to ascertain the presence or absence of crosslinguistic influence in learners' use of the modal auxiliary verb *will*.

Modal verbs have frequently attracted attention from researchers due to the difficulties learners face in reaching nativelike use, both quantitatively and qualitatively. The current paper selects *will* as the focus of analysis because the acquisition of this modal auxiliary is predicted to present different sets of difficulties for the two groups of learners.

2. Literature Review

This section reviews previous studies relevant to the focus of the present study. Section 2.1 considers corpus studies of L2 modal verb use. Section 2.2 examines modal auxiliaries in Chinese and their similarities to the English modal *will*. Section 2.3 combines the conclusions of the two preceding sections to explain the rationale behind the current study and its hypotheses.

2.1 Corpus Studies of L2 English Modal Verb Use

This section provides an overview of corpus-based studies examining the use of modal verbs by Japanese learners of English (JLE) and Chinese learners of English (CLE).

Nakayama (2020) compares JLE and two groups of native speakers (students and teachers) using ICNALE's written component. JLE are found to overuse *can*, *should* and *must*, but underuse *will* and *would*. Nakayama suggests this reflects the greater difficulty of epistemic modality markers, but does not consider learners' proficiency levels. Nakayama (2021), using the spoken module of ICNALE, finds that JLE at A2 and B1 levels underuse *could*, *might*, *would* and *will*, and use modal verbs to express deontic modality more frequently than epistemic modality, contrasting with native speakers. While Nakayama provides analysis for a selection of individual verbs, there is no specific explanation for the underuse of *will*.

Xiao (2017) compares data from learner and native corpora and reports that CLE overuse *must*, *should*, *will* and *can*, but underuse *would*, *might* and *could* in their writing. Likewise, in spoken language, CLE overuse *must*, *should*, *will* and *can*, but underuse *would* and *might*. Xiao adopts an analytical framework from functional grammar, which groups *will* with other "middle-value" modals, *would* and *shall*. As a result, the analysis cannot adequately explain the high frequency of use of *will*.

Yang (2018) reports that modal verbs appear more frequently in learners' academic writing than in published academic papers, and that learners overuse *can*, *will*, *could* and *would*. Yang suggests that one-to-one translations of modal verbs in course books may cause pragmatically inappropriate uses of *should* by CLE (p. 127).

Taken together, the above studies appear to show trends towards underuse and overuse of *will* by JLE and CLE respectively. This paper aims to directly compare the two learner groups using a unified data set and offer explanations for any differences observed. While a principled analysis of course books is beyond the scope of the current paper, the characteristics of learners' native languages will be analyzed as one cause of differing patterns of use of modal verbs in L2 English. In particular, real and perceived similarities between modal verbs in Chinese and English will be shown to influence the use of *will* by CLE. The following section briefly compares and contrasts *will* and corresponding modal verbs in Chinese.

2.2 Modal Auxiliaries in Mandarin Chinese

Mandarin Chinese possesses a wide inventory of modal verbs, but this section will focus on the modal verb *huì* due to its similarities with *will*.

Tsai (2015) distinguishes 5 uses of *huì* as a modal verb. While the future and epistemic uses in (2) and (5) correspond to *will*, what Tsai terms "dispositional" (3) and "generic" (4) modals are not typically expressed using *will* (Tsai's idiomatic English translations of the Chinese sentences have been slightly adjusted). Dispositional and generic modals are referred to below as "non-future" uses. Table 1 summarizes the partial correspondence between *huì* and *will*.

(1) Yiqian waijiaoguan dou hui shuo fayu. [dynamic modal]
Before diplomat all can speak French
'In the past, all diplomats could speak French.'

(2) Waijiaoguan hui changchang lai zheli. [future modal]
Diplomat will often come here
'Diplomats will come here often.'

(3) Waijiaoguan changchang hui lai zheli. [dispositional modal]
diplomat often tend.to come here

'Diplomats often tend to come here.'

(4) Shui hui wang dichu liu. [generic modal]
water HUI towards low.land flow
'Water flows to lower places.'

(5) Waijiaoguan dagai hui lai zheli. [epistemic modal]
diplomat probably Irr come here
'Diplomats will probably come here.' (Tsai, 2015, p. 278)

Table 1. Correspondence between *huì* and *will*

Uses of *huì* (Tsai 2015)	Correspondence with *will*
1. Ability	no
2. Future	yes
3. Dispositional	limited
4. Generic	limited
5. Epistemic	yes

Examples of uses of *will* resembling the generic and dispositional uses of *huì* are shown in (6) and (7) below. These are examples of corrected learner production displayed on the English Grammar Profile Online and are described there as “habitual and typical” (6) and “willfulness or disapproval” (7) uses of *will*. Such uses are deemed to be limited for the following reasons. First, they are categorized at CEFR C1 and C2 levels respectively, so are typically acquired only at the highest proficiency levels. This is likely related to their low frequency of use by native speakers. Given the high CEFR ratings, it is unlikely that learners at the proficiencies focused on in this study will have received sufficient input to use them in their own writing.

(6) “habitual and typical” (C1)
Can use 'will' to talk about something which is typical or habitual.
Example: She will often knock on the door to see you.
(Japan; C1 EFFECTIVE OPERATIONAL PROFICIENCY; 1993; Japanese; Fail)

(7) "willfulness or disapproval" (C2)
Can use 'will' to talk about general behaviour, often disapprovingly.
Example: Indeed no one can imagine what children will do!
(France; C2 MASTERY; 1993; French; Pass) (English Grammar Profile Online)

Secondly, as alluded to in the description in (7), *will* used to talk about general or typical states of affairs often expresses an air of disapproval which is absent from dispositional and generic uses of *huì*. Carlson (2012, p. 834) discusses a further restriction, namely that habitual *will* cannot appear with individual-level states. Even where a habitual reading is plausible, as in (8c), this reading is excluded in favor of an epistemic reading in which the speaker is making a prediction about future conditions.

(8) a. Bob will be an attorney.
b. The girl will like ice cream.
c. The weather will be very mild here. (Carlson, 2012, p. 834)

2.3 Functional Similarities Between L1 and L2 and the Potential for Crosslinguistic Influence

Section 2.2 demonstrated that *will* and *huì* have limited functional similarities, namely their future and epistemic uses. In contrast, dispositional and generic uses of *will* are infrequent, marked and unlikely to be encountered in input learners receive. Nonetheless, there is potential for crosslinguistic influence in all five uses shown in Table 1 above. According to Jarvis & Pavlenko (2008, pp. 178–180), crosslinguistic influence typically occurs where there are subjective similarities between L1 and L2. A pertinent example is reported in Odlin (2008, pp. 317–318), who refers to a study by Sastry-Kuppa (1995). This study showed that native speakers of Tamil used *will* as a marker of habitual aspect, not only in the present tense but also in the past tense, where *would* or *used to* would be appropriate. Sastry-Kuppa concludes that this reflects overgeneralization of the similarities between *will* and the future tense marker in Tamil.

If CLE overgeneralize the similarities in Table 1 and assume functional equivalence in categories 3 and 4, they are expected to overuse *will* to express dispositional and generic meaning. This in turn may lead to overuse of *will* overall. This indeed appears to be the case, as the corpus studies in section 2.1 have revealed. What

the present study seeks to demonstrate is that CLE are indeed using *will* to express dispositional and generic meaning where native speakers do not (or do so at a significantly lower frequency).

In contrast to CLE, previous studies found that JLE underuse *will* in comparison to native speakers. This can be explained by considering the means of expressing modality in Japanese. Modal verbs such as *can*, *should* and *must*, which JLE were found to overuse, are typically expressed in Japanese using sentence final expressions or verbal inflections (9). In contrast, future and epistemic meanings are not expressed by a dedicated, obligatory morpheme, although *–daro* or *–ka mo shirenai*, expressing a subjective judgement of probability, are optionally attached to the non-past form of the verb (10). Furthermore, dispositional and generic meaning can also be expressed using an unmarked verbal form. This means that Japanese lacks formal equivalents to *will* and therefore L1 forms are not predicted to aid the acquisition of L2 forms. This lack of morphological salience in L1 is expected to manifest itself in underuse of *will* compared to both native speakers and CLE.

(9) Gakusei wa apuri de benkyō {suru koto ga dekiru / suru beki da / shinakereba naranai}.[i]
Student TOP app INS study {do NMLZ NOM can / do ought.to COP / do-NEG-COND become-NEG}.
'The students {can / should / have to} study by accessing the online resources.'

(10) Gakusei wa apuri de benkyō suru (darō / ka mo shirenai).
Student TOP app INS study do (COP-CONJEC / Q also know-POTEN-NEG)
'The students (will probably/might) study by accessing the online resources.'

Newbery-Payton and Mochizuki (2020) analyzed L1 to English translations by CLE and JLE in order to explore the effect that the absence or presence of comparable L1 forms has on the production of L2 forms. Errors of omission of *will* appeared exclusively in JLE data, while translations by CLE were characterized by inappropriate use of *will* in habitual senses. Newbery-Payton and Mochizuki explained these contrastive error trends through reference to the kinds of characteristics of Chinese and Japanese discussed above.

3. Research Design

The current paper seeks to verify the findings of Newbery-Payton & Mochizuki (2020) using different methodology. Specifically, it adopts a larger data set, examines a different task format, and compares both native speakers and learners at different proficiency levels using statistical testing.

3.1 Aim and Research Questions

This paper considers the research questions listed below. Research Questions 1 and 2 concern the overall frequency of use of *will*. CLE and JLE are expected to differ in their use of *will*. CLE are expected to exhibit a higher frequency of use than NS, while JLE are expected to exhibit a lower frequency of use. In addition, both groups of learners are expected to become more native-like in terms of frequency of use at higher proficiency levels.

RQ1: To what extent do CLE and JLE differ in their use of the modal auxiliary verb *will*?

RQ2: To what extent does the use of *will* by CLE and JLE become more native-like with increasing proficiency?

Research Question 3 concerns the effect of L1 forms on the use of *will* in L2 English. CLE are expected to overuse *will* in dispositional and generic senses, as a result of overgeneralization from L1. A similar phenomenon is not expected in the JLE data due to the lack of functional equivalents in L1.

RQ3: To what extent can the use of *will* by CLE be explained by reference to L1 forms?

3.2 Data and Method

Data is sourced from the Written Essays module of the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2013). Use of this data set provides the following advantages. First, two essay topics are specified for participants to write about, allowing both topic control and comparison of topics. The prompts for

each essay are shown in (11) below. In the remainder of this paper, topics A and B are abbreviated as “PTJ” (part-time job) and “SMK” (smoking) respectively.

(11) Do you agree or disagree with the following statements? Use reasons and specific details to support your opinion.
(Topic A) It is important for college students to have a part-time job.
(Topic B) Smoking should be completely banned at all the restaurants in the country. (Ishikawa, 2013, p. 97)

Although neither topic uses *will*, the prompt for the SMK task includes the modal verb *should*, which potentially affects the use of other modal verbs. Nevertheless, use of this data set is still preferable to the translation task used by Newbery-Payton & Mochizuki (2020), as the latter task type may enhance the potential for L1-related effects to occur. This is because learners may be directly influenced by features of the L1 text they are required to translate. Further discussion of task-related effects is provided in section 5.

The second advantage of using ICNALE is that it includes data from learners judged to be at A2, B1 and B2 CEFR levels. This allows pseudo-longitudinal analysis in order to examine the effect of proficiency. As alluded to in the previous section, crosslinguistic influence is predicted to be mediated by increasing proficiency.

48 essays on each topic were randomly selected from the data sets for JLE and CLE at A2, B1-1 and B1-2 levels. This reflects the size of the smallest of the subcorpora under consideration (JLE B1-2, N=49). B2 level learners were excluded from the analysis due to data size limitations. ICNALE contains three groups of NS; the student group was selected for analysis as this was considered to best match social characteristics of the learner groups. A total of 672 files totaling 154,088 words were selected for analysis. Summaries of the data size and learner attributes are provided in Tables 2 and 3.

Table 2. Data Summary

	PTJ					SMK			
	A2	B1-1	B1-2	Total		A2	B1-1	B1-2	Total
CLE	10923	11972	12287	35182	CLE	10713	11203	11438	33354
JLE	10933	10540	11057	32530	JLE	10423	10349	10850	31622
Total	21856	22512	23344	67712	Total	21136	21552	22288	64976
NS		10774		78486	NS		10626		75602

Total Files: 672 / Total Words: 154,088

Table 3. Summary of Learner Attributes

	CLE				JLE		
	A2	B1-1	B1-2		A2	B1-1	B1-2
F	25	24	21	F	22	17	21
M	23	24	27	M	26	31	27
Average age	19,2	19,5	19	Average age	18,5	18,6	18,8

Data was tagged using TagAnt and relevant examples were then extracted using AntConc. Each instance of *will* was examined within the wider context of the essay to determine the most likely intended meaning. In particular, “non-future” uses of *will* were identified and extracted for further analysis (see section 4.2).

4. Results

Quantitative analysis, relating to RQ1 and RQ2, is presented in section 4.1. Qualitative analysis, relating to RQ3, is presented in section 4.2.

4.1 Quantitative Analysis

Table 4 shows the adjusted frequency of use of *will* by each group of learners in the two tasks. The data from the SMK and PTJ tasks are also shown in Figures 1 and 2 respectively. Black dotted lines in the figures show the performance of NS on each task.

Table 4. Adjusted Frequency (per 10,000 words) of *will*

	Part Time Job				**Smoking**		
	A2	B1-1	B1-2		A2	B1-1	B1-2
CLE	67.75	78.5	74.9	CLE	104.5	80.3	62.9
JLE	47.6	58.8	44.3	JLE	23.0	32.9	37.8
NS		72.4		NS		54.6	

On both tasks and at all proficiency levels, adjusted frequency is highest for CLE and lowest for JLE, with frequencies for NS appearing between the two groups of learners (the exception is the PTJ task, where adjusted frequency is slightly higher for NS than for CLE at A2 level). However, Figures 1 and 2 reveal different trends beyond these general similarities.

On the SMK task, frequency is particularly high for A2 CLE and particularly low for A2 JLE, resulting in a high degree of disparity between the two groups at A2 level. With increasing proficiency, however, CLE frequency of use falls and JLE frequency of use rises. In this way, proficiency effects are visible, with both groups of learners approaching native-like frequencies of use at higher proficiency levels.

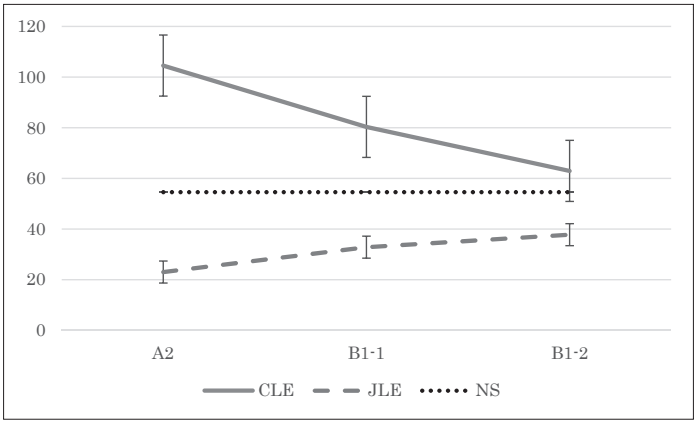


Figure 1. Adjusted Frequency (per 10,000 words) of *will* in "Smoking" Task

The PTJ task displays less variation, both between groups and over different proficiency levels. In both groups of learners, there are marginal increases in frequency at B1-1 level, followed by marginal decreases in frequency at B1-2 level. Furthermore, A2 level learners' frequency of use is already relatively close to that of NS. As a result, there are no obvious proficiency effects comparable to those in the SMK task.

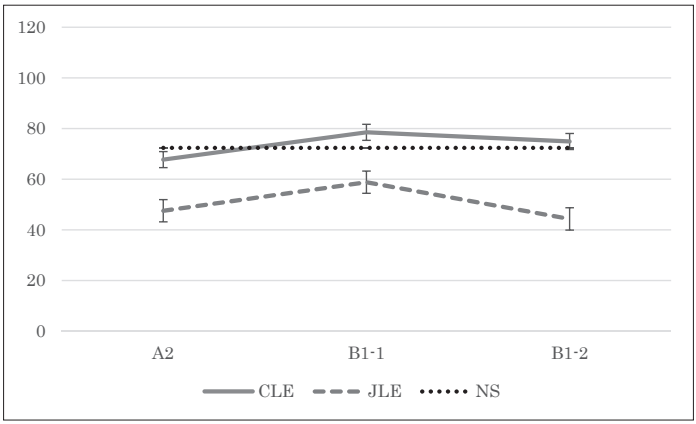


Figure 2. Adjusted Frequency (per 10,000 words) of *will* in "Part-Time Job" Task

The data for each task was next tested for statistical significance. A Kruskal-Wallis test (df=6, χ^2 =57.334230, p= 1.563306e-10) revealed that the groups were not homogenous on the SMK task. Results of post-hoc tests (Dunn method, adjusted with Holm FWER for multiple comparisons) are reported in Table 5. A Kruskal-Wallis test on the PTJ data found no significant difference between groups.

Table 5. Dunn Adjusted p-values for Pairwise Comparisons (SMK)

	CLE_A2	CLE_B1-1	CLE_B1-2	JLE_A2	JLE_B1-1	JLE_B1-2
CLE_B1-1	.289					
CLE_B1-2	**.035**	1				
JLE_A2	**<.001**	**<.001**	.014			
JLE_B1-1	**<.001**	**.012**	.121	1		
JLE_B1-2	**<.001**	.052	.364	1	1	
NS	**.019**	1	1	**.027**	.207	.537

The significant differences in Table 5 (shown in bold) can be summarized as follows. CLE at A2 level use *will* significantly more frequently than almost all other groups, the only exception being B1-1 level CLE. CLE at B1-1 level also show significantly higher frequency of use than A2 and B1-1 level JLE. No significant differences were found between learners at B1-2 level. Taken together, the results corroborate the trends observed in Figure 1.

Comparing learners to NS, while A2 level learners show significantly higher (CLE) or lower (JLE) frequency than NS, at B1-1 and B-1-2 these differences are no longer significant. In other words, non-nativelike frequency of use, both overuse and underuse, is limited to A2 level. This provides answers to RQ1 and RQ2: CLE and JLE differ significantly at A2 and partially at B1-1 level; from B1-1 level onwards, learners' frequency of use converges and becomes more native-like.

These results are not simply a reflection of idiosyncratic modal auxiliary use by a minority of learners. As summarized in Table 6, the percentages of learners in each group using *will* on at least one occasion in their writing show largely similar trends to the adjusted frequencies shown in Table 4. In other words, the high but falling frequency of use by CLE and the low but rising frequency of use by JLE appear to be characteristics of each group as a whole.

Table 6. Percentage of Learners Using *will* in the SMK Task

	A2	**B1-1**	**B1-2**
CLE	83	79	65
JLE	33	40	44

The next section considers RQ3, namely whether the significantly higher frequency of use of *will* produced by CLE can be explained, at least in part, by crosslinguistic influence in the form of overgeneralization of L1 forms.

4.2 Qualitative Analysis

RQ3 concerns the extent to which the "non-future" uses of the Chinese modal auxiliary *huì* might influence CLE use of *will*. In order to answer this question, uses of *will* were categorized and the proportion of "non-future" uses calculated. Analysis was

conducted by the author using the following heuristics (judgement by multiple annotators was not feasible due to practical constraints). Instances of *will* were considered "non-future" if they expressed states of affairs current to the reference time and could be replaced with present tense forms with minimal change of meaning. This amounts to a division between the future and epistemic uses of *will* on the one hand, and generic and dispositional uses on the other. This level of granularity was judged to be sufficient for the purposes of the current research question as it reflects the key parallels between English and Chinese. Conditional sentences containing *if* such as (12) allow an epistemic interpretation and so were not considered "non-future" uses. Sentences including *when* were judged on the content of the sentence and the wider context of the essay. For instance, (13) was regarded as expressing a future state of affairs whereas (14) was regarded as expressing a generic state of affairs (note the use of *sometimes*); only the latter was considered a "non-future" use.

(12) If the law of banning is through, the atmosphere of restaurants **will** be more perfect. CHN_SMK_039_A2

(13) When smokers cut down the number of cigarettes, the good dining atmosphere **will** be built easily. CHN_SMK_269_B1_1

(14) They said sometimes inspire **will** come across in their mind when they smoked. CHN_SMK_310_A2

Contrary to expectations, A2 level JLE also showed some non-future uses of *will* (15). However, as proficiency rises, non-future uses of *will* largely disappear from the JLE data, while continuing to account for 15-20% of the overall use of *will* by CLE (Table 7). Examples of "non-future" uses of *will* by B1-level CLE are given in (17) and (18). The persistence of such examples suggests that CLE continue to use *will* in a manner analogous to L1, providing partial confirmation of the prediction for RQ 3.

(15) Especially, in the restaurant, many people **will** enjoy eating and talking with friends or families. JPN_SMK_344_A2

(16) There are many people who like smoking, even in the public places they **will** take a cigarette in hand. CHN_SMK_289_A2

(17) Finally, it is a good idea to ban smoking in any restaurants because When someone smoke cigarettes, harmful gases **will** arise and fulfill the room. CHN_SMK_104_B1-1

(18) France has forbidden people smoking in cafes a few years ago. The citizens who defy it **will** be punished and feed for a lot. CHN_SMK_363_B1-2

Table 7. Frequency and proportion of non-future uses of *will*

	A2	A2 (%)	B1-1	B1-1 (%)	B1-2	B1-2 (%)
CLE	20	18	18	20	11	15
JLE	5	21	3	9	1	2

It must be recognized that "non-future" use of *will* alone cannot explain the significant differences in frequency at A2 and B1-1 levels. The phenomenon might best be understood as one expression of L1 transfer occurring more generally in the writing of CLE. The remainder of this section will consider one further aspect of the low frequency of use by JLE in the SMK task, namely errors of omission.

JLE are expected to omit *will* in obligatory contexts more frequently than CLE due to the absence of functional equivalents to *will* in Japanese. One such situation is in conditional clauses, as verbs in consequent clauses are frequently marked with *huì* in Chinese but receive no dedicated morphological marking in Japanese. Conditional clauses were extracted from the data set by searching for sentences including *if* then filtering manually. Examples are provided below, with relevant errors in bold. As predicted, JLE show a greater raw frequency and proportion of errors of omission in obligatory contexts (Table 8). Extraction of all errors of omission was beyond the scope of the current paper, but it seems reasonable to expect errors of omission to occur more frequently throughout JLE's writing, not only in conditional clauses.

(19) However, if smoking is banned at all the restaurants, the smoker **is** uncomfortable

and dissatisfied. W_JPN_SMK0_312_A2

(20) If smoking is forbidden at all restaurants, I think only nonsmokers and light smokers **enjoy** meals. W_JPN_SMK0_269_B1_1

(21) If smokers stop smoking at the public places in order not to be finned, the public places **become** more comfortable and cleaner. W_JPN_SMK0_179_B1_2

Table 8. Omission of *will* in Obligatory Contexts (Conditional Clauses)

	A2	A2 (%)	B1-1	B1-1 (%)	B1-2	B1-2 (%)
CLE	3	4	3	6	2	5
JLE	10	14	12	17	13	17

In summary, it appears that crosslinguistic influence may have affected the overall frequency of use by the two groups of learners, as well as influencing the proportion of uses of will expressing “non-future” meaning and the proportion of errors of omission in obligatory contexts. The implications of these findings are discussed in the following sections.

5. Discussion

This section considers a number of issues raised by the present study and their implications for future research. Section 5.1 reviews the study’s main findings and their relation to theoretical distinctions in the field of second language acquisition. Section 5.2 considers task- and proficiency-related effects in relation to previous studies, while Section 5.3 considers task-related and other effects within the current data set.

5.1 Crosslinguistic Influence, Suppletion and Addition in Second Language Acquisition

While L1-related effects on acquisition (measured in terms of accuracy of use) have been demonstrated for a range of grammatical categories (Luk & Shirai, 2009; Murakami & Alexopoulou, 2016), few such effects have been demonstrated for *will* or

for modal verbs more generally. Recent explanations offered for the overuse or underuse of individual modal verbs (see section 2.1) rest on categorizations of modal verbs as a group, without considering L1-related factors. This study's findings suggest that specific consideration of relevant L1-related factors can provide nuance that such approaches miss.

Murakami & Alexopoulou (2016, p. 368) hypothesize that "lack of the equivalent feature in the L1 leads to low accuracy". The present study confirms this for JLE, while also showing that the presence of an (apparently) equivalent feature leads to lower accuracy, due to overuse of the target form. Gabriele (2009) argues that addition (the acquisition of new interpretations of a given linguistic form) should be distinguished from preemption (the ruling out of interpretations present in L1 but not in L2). Gabriele examines the acquisition of different interpretations of progressive forms in English and Japanese and concludes that preemption is more difficult than addition, especially in the absence of explicit input showing otherwise.

In the context of the present study, CLE must preempt the "non-future" uses of *will*, while JLE must acquire the semantics of *will* due to a lack of a functional equivalent in Japanese. If acquisition is incomplete, CLE are expected to overuse non-future senses of *will* and JLE are expected to underuse and/or omit *will* in obligatory contexts. While this is what the results have indicated, the current study does not offer evidence either for or against the assertion that preemption is more problematic than addition.

5.2 Task- and Proficiency-Related Effects in Relation to Previous Studies

Newbery-Payton & Mochizuki (2020) analyzed L1 to English translations by high proficiency learners. The present study, however, found that CLE and JLE had converged by B1-2 level, suggesting that task type influences the extent to which L1-influence occurs. Translation tasks, which provide an L1 text to translate into L2, may induce even higher proficiency learners to emulate certain features of L1 in their L2 writing, whereas free-writing tasks appear to show L1-related effects only at lower proficiency levels. This underlines the importance of confirming research findings using different data sets, task types and groups of learners. In the present study, there was no L1 source text that might induce L1-like norms in L2 writing, such as the inclusion of *will* wherever *huì* appeared in the source text. Furthermore, the essay

prompts did not contain *will*, so direct linguistic influence from the prompts cannot be assumed.

In addition, the studies referred to in Section 2 examining L2 English modal use cannot be said to have fully considered proficiency effects. It is possible that the modal verbs reported to be underused or overused were in fact used at native-like frequencies by higher proficiency learners. Clarification of such issues could be beneficial when considering, for example, which aspects of modal verb pedagogy would most benefit from reconsideration and at which stages of EFL study.

5.3 Task-Related and Other Effects in the Present Study

As stated in Section 4.1, differences between the three groups in the PTJ task were not statistically significant. This may reflect lesser (perceived) "opportunity of use", a term used to refer to "the opportunity the learner is afforded to use a linguistic feature", which can be affected by factors including task type, task topic and document length (Buttery & Caines, 2018, p. 6). Of these three factors, task topic is the most relevant for the ICNALE data.

While neither essay topic discourages use of *will*, the SMK task is arguably more conducive to writing about hypothetical future events, as learners are encouraged to write about the implications of a possible future change in the law. This provides two contexts – sentences including future time reference and consequent clauses in conditional sentences – where functional similarities between *huì* and *will* encourage the use of the latter by CLE. However, while CLE at A2 and B1-1 level did indeed use *will* more frequently on the SMK task, the opposite is true for the B1-2 group. NS and JLE at all proficiencies similarly displayed a higher adjusted frequency on the PTJ task than they did on the SMK task (Table 4).

It is difficult to provide conclusive answers to this puzzling phenomenon, but one explanation may lie in the use of other modal verbs, which are in syntactic competition to appear before the main verb in a given sentence. As stated in section 2, Nakayama (2020) reported overuse of *can*, *should* and *must* by JLE in the ICNALE data. This is likely related to the fact that *should* appears in the essay prompt for the SMK task (11). JLE may have selected *should* more frequently in the PTJ task, leaving fewer opportunities for the use of *will*. If the conclusions of this paper are valid, then CLE are already primed to use *will* due to L1-related factors, causing the prominent differences

between the two groups on the SMK task. NS may be less likely to be influenced by the prompt, given the wider range of linguistic devices available to them.

The PTJ essay prompt does not include any modal verbs so learners are not explicitly induced to select one modal over another and the topic may be more conducive to a mix of temporal references and real and hypothetical situations. This may be why the PTJ task exhibited more homogeneous use of *will* by the three groups.

A reviewer suggests that cultural differences may influence the trends reported in this study. While it is possible that one group of learners is more likely to hedge statements using other modal verbs, thus avoiding *will*, in the view of the author this is more likely to occur with auxiliaries used primarily as deontic modals, which more directly reflect the writer's stance. Chen and Zhang (2017, pp.19–21), in their study of hedging by Chinese and Anglophone writers, report that the only modal verb with significant differences in frequency was *should*; Chinese writers were found to overuse *should* as a deontic modal but underuse it as an epistemic modal. In regard to writing by Japanese native speakers, Takimoto (2015, pp. 95–96) states that although Japanese speakers may express themselves indirectly in their native language, such L1 norms are not necessarily replicated in L2 English writing. Takimoto reports that JLE use boosters (including *will*) as frequently as NS, and hedges significantly less frequently than NS.

High frequency hedges in Takimoto's study include the modal verbs *could* and *may*. A comprehensive study of learners' selection of modal verbs is beyond the scope of the present paper, but the occurrence of these two modal verbs in the current data set can be summarized as follows. First, their frequency generally rises with increasing proficiency for CLE. This might be expected on the SMK task, where the frequency of *will* falls significantly for CLE at B1-2 level (Figure 1). It cannot, however, account for the PTJ task, where the frequency of *will* displays minimal change despite the increase in frequency of these hedging modals.

Second, the frequency of hedging modals is typically lower for JLE than it is for CLE. Furthermore, as learners' proficiency rises, frequency of use either decreases or returns to its original level after an initial rise. The exception is *could* on the PTJ task, where JLE also exhibit a large increase, exceeding the frequency for NS. The general tendency may be another instantiation of underuse of modal verbs by JLE due to the absence of equivalent obligatory morphemes in L1 (see Section 2.3).

It should be noted that the combined frequency of *could* and *may* across tasks and

groups (N=459) is less than half of that of *will* (N=932). In short, it seems unlikely that trends in the overuse and underuse of *will* can be reduced to an epiphenomenon caused by selection trends among other modal auxiliaries.

Use of other modal verbs does not appear to explain the differing frequency of use of *will* by NS either. For instance, adjusted frequencies of *could* and *may* are comparable on both tasks, and the adjusted frequency of *can* is higher on the PTJ task (100) than on the SMK task (59). The PTJ task therefore appears to be generally more conducive to the use of modal verbs – at least for NS. The author hopes to address this issue more fully in future studies.

Finally, another reviewer asks whether the frequency of *will* is related to the frequency of *going to,* particularly at lower proficiency levels. While the highest frequency of use was indeed observed in the data for the PTJ task by A2 level JLE, *going to* appeared only 13 times in the whole data set. Given this low frequency, preference for one future expression over another appears to have a relatively small effect on trends of use, at least for the current topics and task types.

6. Conclusion

This study examined the extent to which CLE and JLE differ in their use of *will* (RQ1), the effect of proficiency on frequency of use (RQ2), and the extent to which overuse or underuse of *will* can be explained with reference to L1 forms (RQ3). Analysis revealed significant differences in the use of *will* by CLE and JLE at lower proficiency levels, whereas learners at higher proficiency levels did not differ significantly from each other or from native speakers. Qualitative analysis showed that non-future uses of *will* were significantly higher among CLE, suggesting learners use the form in an analogous manner to the Chinese modal auxiliary *huì*. JLE do not display this characteristic and also show a tendency to omit *will* in obligatory contexts, suggesting that the absence of a comparable L1 form is one factor in the underuse of *will*.

The current study was limited to written language, so spoken data from ICNALE could also be analyzed in future. Online processing demands during speech may cause higher rates of omission of target forms, particularly at lower proficiency levels. It is unclear, however, whether or not this will significantly affect the differences between

JLE and CLE in terms of frequency of use of *will* or other modal verbs.

While this study has focused on one particular linguistic form, similar methodology could be used to investigate the frequency of other forms. Principled selection of these forms, and of the L1 groups to include in analyses, can be aided by careful consideration of L1 characteristics.

Finally, more attention has been paid in recent years to the interface between the fields of corpus linguistics and second language acquisition (Le Bruyn & Paquot, Eds., 2021). It is hoped that corpus analyses like the present study can complement existing SLA research or provide the impetus for new studies. For example, researchers could examine whether overuse and underuse of modal verbs by different groups of learners are also observable in cloze tasks or other experimental designs.

References

Anthony, L. (2015). TagAnt (Version 1.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Anthony, L. (2020). AntConc (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Binnick. R. (2005). The markers of habitual aspect in English. *Journal of English Linguistics*, *33*(4), 339–369.

Caines, A., & Buttery, P. (2018). The effect of task and topic on opportunity of use in learner corpora. In Brezina, V., & Flowerdew, L. (Eds.). *Learner corpus research: New perspectives and applications* (pp. 5–27). Bloomsbury.

Carlson, G. (2012). Habitual and generic aspect. In Binnick, R. (Ed.). *The Oxford handbook of tense and aspect* (pp. 828–851). Oxford University Press.

Chen, C., & Zhang, L. (2017). An intercultural analysis of the use of hedging by Chinese and Anglophone academic English writers. *Applied Linguistics Review*, *8*(1), 1–34.

Gabriele, A. (2009). Transfer and transition in the SLA of aspect A bidirectional study of learners of English and Japanese. *Studies in Second Language Acquisition*, *31*, 371–402.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. School of Languages & Communication, Kobe University. *Learner Corpus Studies in Asia and the World*, *1*, 91–118.

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. Routledge.

Luk, Z., & Shirai, Y. (2009). Is the acquisition order of grammatical morphemes impervious to L1 knowledge? Evidence from the acquisition of plural -s, articles, and possessive 's. *Language Learning*, *59*, 721–754.

Le Bruyn, B., & Paquot, M. (2021). *Learner corpus research meets second language*

acquisition. Cambridge University Press.

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.

Nakayama, S. (2020). Contrastive interlanguage analysis of modal auxiliary verb usage by Japanese learners of English in argumentative essays. *The IAFOR International Conference on Education – Hawaii 2020 Official Conference Proceedings*.

Nakayama, S. (2021). Modal auxiliary verbs in Japanese EFL learners' conversation: A corpus-based study. *Asia Pacific Journal of Corpus Research*, *2*(1), 23–34.

Newbery-Payton, L., & Mochizuki, K. (2020). L1 influence on use of tense/aspect by Chinese and Japanese learners of English. School of Languages & Communication, Kobe University. *Learner Corpus Studies in Asia and the World*, *4*, 67–93.

Odlin, T. (2008). Conceptual transfer and meaning extensions. In Robinson, P., & Ellis, N. (Eds.). *Handbook of Cognitive Linguistics and Second Language Acquisition* (pp.306-340).

Sastry-Kuppa, S. (1995, March 3). *That's why he will talking for English: The expression of habitual aspect in the English of untutored and low-level tutored Indian speakers*. Presentation at the Ninth International Conference on Pragmatics and Language Learning, University of Illinois.

Takimoto, M. (2015). Assertions and lexical invisibility in EFL learners' academic essays. *Journal of Pragmatics*, *89*, 85–99.

Tsai, W. (2015). On the topography of Chinese modals. In Shlonsky, U. (Ed.). *Beyond functional sequence* (pp.275–294). Oxford University Press.

Xiao, Y. (2017). Chinese EFL learners' acquisition of modal verbs: A corpus-based study. *International Journal of English Linguistics*, *7*(6), 164–170.

Yang, X. (2018). A corpus-based study of modal verbs in Chinese learners' academic writing. *English Language Teaching*, *11*(2), 122–130.

This research has made use of the English Grammar Profile. This resource is based on extensive research using the Cambridge Learner Corpus and is part of the English Profile programme, which aims to provide evidence about language use that helps to produce better language teaching materials. See http://www.englishprofile.org for more information.

[i]Notation follows conventions in the Handbooks of Japanese Language and Linguistics published by the National Institute for Japanese Language and Linguistics.

COND	conditional	INS	instrumental	POTEN	potential
CONJEC	conjectural	NEG	negative	Q	question marker
COP	copula	NMLZ	nominalizer	TOP	topic

（Laurence NEWBERY-PAYTON 東京外国語大学　Email: laurencecnp-tufs@tufs.ac.jp）

「論文」

Developing Classroom Corpus Tagger for Teachers' Reflective Practice: A Spoken Language Tagger to Compile Classroom Corpora

Yukiko OHASHI, Noriaki KATAGIRI, and Takao OSHIKIRI

Abstract

This study presents a browser-based Classroom Corpus Tagger (CCT): Discourse Tagging Assistant. The CCT tool has been developed to markup speaker tags by clicking the mouse and instantaneously encoding language-use tags for compiling a classroom corpus. Classroom corpus compilation involves attaching tags to each transcribed utterance according to the tagging design. Annotation of the utterances by teachers and students requires multi-layered tags to be attached; this is a time-consuming process and sometimes leads to unexpected human errors. Hence, this study attempted to develop a basic discourse tagging assistant, the CCT, to smoothly attach pre-designed tags to each transcribed utterance, requiring less time and energy than manual tagging of transcripts. A case study was conducted to test the validity as well as the availability of the CCT tool. The results revealed that tagging using the CCT helped overcome the complexities related to manual tagging of transcripts. Moreover, using the CCT reduced the tagging time for the transcribers, as compared to manual tagging which was sometimes erroneous. The application of CCT is likely to lessen the workload of building classroom corpora, and eventually, promote classroom-related research by facilitating reflective practices. This study introduces how we created the CCT and displays an example of how we utilize classroom corpora. Accumulating classroom corpora using the CCT will enhance the opportunities for teachers' reflective practice as well as evidence-based foreign and second language classroom discourse analyses.

1. Introduction

Corpus data collected in the language classroom provide evidence for reflective

practice as a means of developing language teachers' skills. One of the distinctive advantages of reflective practice through corpus compilation is that it enables teachers to thoroughly analyze their statements, including the detailed vocabulary they use. Corpus data reveal the types and tokens of each lexical item, facilitating teachers' awareness of vocabulary usage. Comparing those types or tokens with the general vocabulary list, such as the General Service List and the recently presented CEFR-J wordlist[1)], teachers get acquainted with the vocabulary items that they should use intentionally and to the maximum. Drawing on the specifics of teacher talk, questioning strategies, types of feedback, and wait-time pauses revealed from classroom corpora can also raise the teaching awareness of novice teachers. Transcripts from corpus data can facilitate close qualitative inspections to hone teachers' decision-making skills about what they say in the classroom (O'Keeffe, McCarthy, & Carter, 2007). As Walsh (2013) states, reflections through observation of transcribed data are likely to result in an ongoing process of enhanced awareness, training second language teachers to improve their verbal expressions and enhance their knowledge of the interactional processes. From the perspective of teachers' reflective practices, a compilation of classroom corpora aids in language teachers' development, as the ephemeral spoken classroom discourse is made visible.

While the compilation of corpora provides a variety of research sources, it also involves demanding manual work of attaching tags to categorize the classroom's discourse data. Attaching different tags manually according to the quality of utterances requires considerable time, and the tasks are prone to tagging errors, which hinders the research due to the painstaking correction process. For example, manual tagging of each utterance to compile a classroom corpus in the study conducted by Katagiri and Ohashi (2017) was time-consuming; hence, it took more time to start the research work than they had planned. A spoken corpus developed by Katagiri and Ohashi (2017) served as a teacher training tool to compare the quality of six classes of preservice teachers. While tagged corpora could provide abundant research sources, completing one corpus takes a significant amount of time and energy, delaying examination of the researchers or teachers' utilization of the original corpus. To address this issue, a corpus compiling tool was designed that allowed instantaneous attachment of speaker and language tags during the research process.

2. Literature Review on Corpus Annotation Structure

Since the onset of corpus-based research, many corpora have been accessible online. Along with the different types of corpora, a variety of tools for corpus research are being downloaded or accessed online. For example, corpus annotation in text tagging is made possible by tagger tools with embedded digital dictionaries, such as Biber Tagger (Biber, 2010) and TagAnt (Anthony, 2015). Tagged text files allow quick calculation of vocabulary frequencies and grammatical features, leading to fast processing of the text data which can be utilized for original research. Considering the compilation of spoken corpora, the first step is identifying each utterance by a speaker and attaching speaker tags manually. There have been few computer tools that can instantaneously identify the speaker and the languages of the transcribed utterances in the texts. This is why compilers of spoken corpora, including classroom corpora, take a lot of time to complete an annotated corpus.

Classroom interactions taking place in a second or foreign language class (ESL/EFL) follow a hierarchical structure (Sinclair & Coulthard, 1975). The hierarchical classroom structure—comprising the teacher's utterances and the student's responses—can be represented by Extensible Markup Language (XML). XML allows users to define a machine-readable set of rules for encoding documents. XML also enables the users to create original markup frameworks, describing documents that conform to a hierarchical structure because their lower elements (child nodes) are nested in the upper elements (parental nodes). Katagiri and Kawai (2016) designed an XML schema for showing classroom discourse visually through eXtensible Style Language Transformations (XSLT). Compilation of classroom corpora using the XML format allows search of the required data through XSL. This provides further improved processing in XPath (refer to Katagiri & Kawai [2016] for details regarding the XPath and XML Schema). For these reasons, we propose to use the XML form for compiling classroom corpora.

It is the corpus markup that defines the availability of classroom corpus. Corpus markup refers to a system of codes inserted into a document, stored in electronic form, or transcribed texts to provide information about the text (McEnery, Xiao, & Tono, 2006). Referring to McEnery et al. (2006), markup helps to structure information, separating documents into appropriate sections with headings, sub-headings, and

paragraphs. It also enables the inclusion of meta-information collected for the corpus. The need for markup to build a classroom corpus can be summarized in the following three perspectives:

1. Markup allows a broader range of research questions according to researchers' needs;
2. Pauses and paralinguistic features, such as laughter and gestures, can be identified through markup; and
3. Corpus markup parallels the existing linguistic transcription.

Ohashi (2015) arranged twenty-four different types of tags in annotation design to be placed in utterances according to their attributes. For example, she annotated the utterance character, "What's the date today?" as <teacher><eng><question>What's the date today?</question></eng></teacher>, indicating that the speaker represented by "<teacher>," asked a question (annotated as "<question>") in English (by "<eng>"). Besides speaker tags, activity tags attached to each classroom activity enabled the teachers and the researchers to reflect on their utterances to improve their articulation.

Annotation designs vary according to the researchers' needs. For example, Ohashi and Katagiri (2016) attached additional tags to differentiate explicit or implicit explanations, in addition to speaker tags and language tags, to examine teachers' explicit instructional roles. Ohashi and Katagiri (2016) also translated Japanese utterances into English and annotated the translations with translated language (TL) tags, <TL></TL>.

Compiling a classroom corpus involves placing tags that represent the quality of classroom discourse. The annotation process is time-consuming and an arduous task because each corpus requires coders' judgment on the classroom discourse quality and use of target language followed by the coders' manual annotation, besides transcription of the recorded speech. The annotation design depends on the research purpose. The more tags the annotation design requires, the more time it takes to complete a whole classroom corpus. With the aim of addressing this challenge, this study designed an original tool to attach designated tags to the text smoothly to lessen the burden of manually compiling a classroom corpus.

According to Walsh (2013), and Mann and Walsh (2017), language teachers'

reflective practice occupies a central position and is of considerable significance in professional education. Reflective practice has been conceptualized differently, and no commonly agreed definition exists. The definitions vary with respect to the extent to which the class focuses on interaction or action (Mann & Walsh, 2017). Some reflective practices emphasize the exploration of experiences that lead to new understandings through engagement in repairs and review (e.g., Boud & Walker, 1998; Zeichner & Liston, 1996), while others highlight critical self-awareness (e.g., Bailin et al., 1999). Both quantified and qualified data from classroom corpora can be used as tools for language teachers to conduct reflective practices, the importance of which has been established in teacher training.

One of the influential reflective practice models is the phased steps summarized by Zwozdiak-Myers (2012). The steps are as follows: 1) observations and reflections; 2) abstraction and conceptualization that produce new understanding; and 3) active experimentation, in which reflection turns into repairs, or improved teacher talk. Mann and Walsh (2017) argue that reflective practices are conducted in stages and phases in which novice teachers analyze and evaluate their classes, to make them more effective.

Teacher education literature describes reflection as an essential aspect of professional practice (e.g., Harkin, 2005; Pollard, 2005; Alger, 2006). Considering the object of classroom corpus compilation is to review and develop the current classes to be better; corpus creation facilitates teachers' reflective practices. Reflective practice enables teachers to observe their performance from a socio-cultural perspective, where learners interact with experts, leading them to better understanding (Walsh, 2013). The corpus-based studies of Ohashi and Katagiri (2016) and Katagiri and Ohashi (2018) revealed the effects of social roles involving scaffolding in the classroom, as explained by Vygotsky (1978). They argue that the insights obtained through reflective practices combined with compiled corpora contribute to teachers' training and professional development. The outcomes of compiling a classroom corpus are likely to contribute to the accumulation of recorded classroom spoken data and provide evidence for reflective practice. Integrating corpus data yielding outcomes pertaining to the reflections is likely to assist teaching professionals in gaining a new understanding of their socio-cultural roles. This study aims to develop a classroom corpus compilation tool to help language teachers compile their original corpus that can be used by them for reflective practice.

3. Classroom Corpus Tagger

3.1 Classroom Corpus Tagging Structure

The classroom corpora start with transcription of teacher and student utterances in the classroom. For transcription, meta-information is required to yield details of what was said, by whom, and in which language. Then, to start with, the utterances and interactions between the teacher and students, or among the students in pairs or groups, are transcribed. Thus, the transcripts require the speaker and language tags as their founding information in the classroom discourse hierarchy. The bottom hierarchical rank can extend to higher ranks, according to the discourse quality of the utterances (see Sinclair & Coulthard [1975] for the detailed classroom hierarchical structure). The main tagging of the classroom corpus utilizes the hierarchical foundation that deals with the speaker and the usage of language.

Figure 1 illustrates the hierarchical structural design of the tagged classroom corpus in XML, using a short transcript line, “Hello.” The utterance (i.e., the transcript line “Hello.”) has a start tag and an end tag; in this case, <English></English>. The language tag implies the speaker’s utterance, representing itself, written generically as <speaker></speaker>.

```
<root>
    <body>
      <.....>
        <speaker>
          <English>Hello.</English>
        </speaker>
      </.....>
    </body>
</root>
```

Figure 1. Classroom corpus tagging structure in XML caption

The speaker tags can represent other speaker types, such as homeroom teachers, students, and assistant language teachers. Likewise, the language tags can include other specific language names, such as Japanese, depending on the classroom context. The upper ranks in the hierarchy (depicted as “<.....>” in Figure 1) can have expanded rank names, based on the users’ or researchers’ interests and needs. Some examples are

language activity and language skills, such as speaking and reading, interaction types. Thus, the rank expansion can describe the classroom teaching exchanges as needed.

3.2 Procedures of Developing the Instantaneous Annotation Tool

This section describes the design of the Classroom Corpus Tagger (CCT)[2] Version 1.0 developed in this study. The CCT enables tagging of both speaker and language tags instantaneously with utterances made in the classroom, which significantly reduces the tagging time. Speaker tags are determined instantaneously by the CCT. Language tags can be easily selected by simply clicking on the utterance result line. For developing the language tags, we took advantage of the fact that English is a single-byte character and Japanese is a double-byte character. Moreover, if it is a single-byte character, the tag will be surrounded by <eng></eng>. Similarly, if it is a 2-byte character, it is tagged with <j></j>. In addition, if a sentence contains a mixture of 1-byte and 2-byte characters, <mix></mix> is added outside the <eng></eng> and <j></j> tags. This makes it possible to generate language tags as soon as the characters are entered.

As for speaker tags, there is no unified standard, and language researchers have been assigning them arbitrarily. To cater to this, CCT allows users to freely set their own speaker tags. When the user clicks on the conversation result line, the speaker tags are switched in the order that the user has set in advance. This reduces the burden of tag input for the user.

The CCT can be operated from a personal computer and potential human errors, such as forgetting to enter the closing tag during manual tagging, can be reduced. The current CCT version is downloadable and can be activated in both Windows operating system and macOS. Thus, the CCT operates offline and does not require an internet connection. The programming language used in CCT is JavaScript, and it works on a browser compatible with Google Chrome, Firefox, and Edge (Internet Explorer is not recommended). CCT users just unzip the software file, index.zip, and display the index.html in an HTML browser to activate the CCT program. The CCT startup screen will appear in Google Chrome, one of the recommended HTML browsers that can run on both Windows and Mac. Figure 2 depicts the elements of CCT.

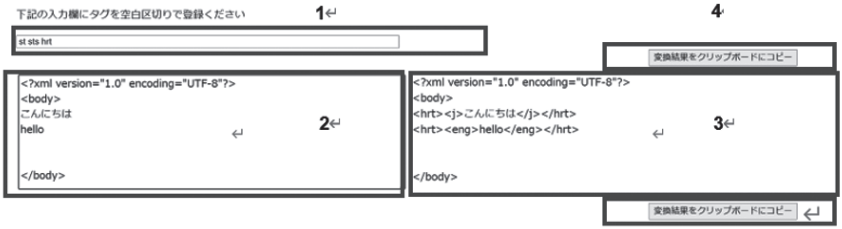


Figure 2. CCT startup screen configuration

Note: Box 1 is speaker tag input box where the Japanese instruction above says, "Please register tag names in the box below spaces." The default values are *st*, *sts*, and *hrt*, meaning st = student, sts = students, and hrt = homeroom teacher. Box 2 is transcription input box. Box 3 is tagged transcription space. Boxes 4 are "Copy conversion results in clipboard buttons" (placed at the top and the bottom of the tagged transcription space, box 3 in Figure 2).

The elements of the CCT are:

1. Speaker tag input box

The speaker tag input box 1 (Figure 2) indicates types of the speaker tags, represented by "hrt", "st", "sts", and "ALT". Users can arbitrarily assign speaker tags according to their needs. For example, if users need to distinguish between individual students —for example between "st1" and "st2" instead of just "st"—they can type in the new tag names in the following manner: The speaker tag can accommodate a maximum of 300 characters, including white spaces, and there is no limit to tag variation. For example, if the tag name is five characters in length, there can be up to 50 different tags (with 49 white spaces in between each name), and if the name contains two characters, there can be up to 100 tags (with 99 white spaces).

2. Transcription input box

The transcription input box 2 (Figure 2) is for entering the transcripts. By entering the text in the area between <body> and </ body>, one can convert it to an XML format. The instantaneously tagged text will appear in box 3 (Figure 2). You can copy and paste the transcribed text created in advance, or you may type in your text directly. Once you start clicking on the XML tagged line in box 3, the tagged lines are fixed, the original

transcript in box 2 will be fixed, and boxes 2 and 3 will not accept any transcript change (i.e., insertion, correction, or deletion). Clicking on box 2 will reset all the speaker tags. We may consider this mechanism to be one of the limitations we need to address for further modification. In this regard, we need to carefully insert the text in box 2. You can enter text of more than 1,000 lines. Additionally, Firefox is faster than Google Chrome, as the transcript lines get incremented due to the browser characteristics.

3. Instantaneous generation of XML tags

When a character is entered in the local text input box, tagged XML is instantaneously generated on the right side of the screen. Figure 3 demonstrates a sample. The upper left box displays the default values of the speaker tag names, st, sts, and hrt. If you enter "Hello." in the lower-left box, the right-hand area will display the tagging result, <hrt> <eng> Hello </eng> </hrt> immediately. The outermost tag, <hrt> is from the final default speaker tag name, hrt. Clicking on the conversion result line cyclically switches to hrt → st → sts → hrt, and so on. When you need to use <st> instead of <hrt>, click on the tagged line, and, you will have the next <st> tag. Similarly, if you want to make it <sts>, click on the same line, and you will replace the current tag, <st> with <sts>. Thus, it is possible to select the desired tags simply by clicking without typing the tag every time you need it.

Language tags are also placed instantaneously according to the entered text (Figure 3). English utterances are tagged with <eng> </eng>, while Japanese utterances are tagged with <j> </j>. If the line includes both Japanese and English utterances, <mix> </mix> will be inserted. For example, entering "Hello" in the transcription input box (the lower-left box in Figure 3), the right-hand area will display "<hrt> <eng> Hello </eng> </hrt>." Thus, the mixed language text with English and Japanese, for example, "Hello こんにちは" turns into <hrt> <mix> <eng> Hello </eng> <j> こんにちは </j> </mix> </hrt >.

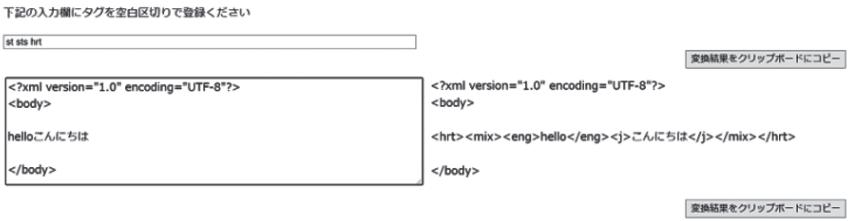


Figure 3. Tagged text displayed in XML tag generation space for “helloこんにちは”

4. Copy tagging results to the clipboard

After feeding the transcript in box 2 followed by adjusting the speaker tags in box 3 (Figure 2), we can copy and paste the tagged transcripts to edit and compile the full-text classroom corpus in the following manner:

1) Press the “copy conversion result to clipboard” button (placed both on the upper and lower right) as indicated by box 4 (Figure 2), to retain the copied information.
2) Paste the copied information into a text editor.

4. Pilot Study: Compiling Classroom Corpora Using the CCT

This section discusses the pilot study conducted to test the CCT’s reliability as a tagger by compiling a mini classroom corpus. The study aims to examine 1) whether the number of vocabulary items in the tagged transcripts (i.e., transcribed classroom utterances tagged by the CCT) is identical to those contained in the manually tagged transcripts, and 2) how the tagging outcomes of the CCT can be utilized.

4.1 Materials and Methods

Five student teachers (four juniors and one senior) at a national university in Japan participated in the pilot test. They signed a consent form showing their willingness to share their English lesson transcripts for the purpose of testing the precision of the tagging process by the CCT. Figure 4 shows the test procedure. The

first two steps involved manual tagging of the classroom transcripts (Step 1) and instantaneous tagging of the same transcripts with the CCT (Step 2). Figure 5 illustrates the tagged samples.

1. Tag transcripts manually in XML format; tag set: (speaker; t, s, ss/ language; mix, eng, j)
2. Tag the same transcripts by the Classroom Corpus Tagger (CCT)
3. Correct tagging errors with an XML document editing software.
4. Extract instructor (student teacher) English utterances using XSL transformation (XSLT).
5. Compare the CCT extraction and the manual instruction.
6. Survey the participant reflections.

Figure 4. Procedure of testing manual and the classroom corpus tagger (CCT) tagging

Raw transcripts [English translation]	➡	Tagged transcripts
-Single language use interaction by the teacher and the student		-Single language use interaction by the teacher and the student
T: そういうとき何するためのもの、じゃあナイフって。[So, what is a knife for?]	⇨	<t><j>そういうとき何するためのもの、じゃあナイフって </j></t>
S: 切る [To cut things.] T: うん、切る。 [Yes, to cut things.]	⇨	<s><j> 切る </j></s> <t><j> うん、切る </j></t>
-Mixed language use by the teacher T: だから [So,] something to cut.	⇨	-Mixed language use by the teacher <t><mix><j> だから </j><eng> something to cut </eng> </mix></t>

Figure 5. Tagging samples

Following the first two steps for tagging classroom transcripts, the tagging errors detected by Editix[3)], an XML document editing software, were revealed in the third step. The manual tagging (Step 1; Figure 4) was prone to XML grammar errors, while the CCT tagging (Step 3; Figure 4) resulted in no such errors. Table 1 illustrates the summary of errors that appeared during manual tagging. The manual tagging resulted in more end tag errors than start tag errors, except for the start tag, *eng* in the transcript M2 for a good reason. Manual tagging is likely to cause errors, and Table 1 suggests that human coders are more likely to miss end tags due to inattention to attaching an end tag, which does not occur in the CCT tagging.

Table 1. Error summary of manual tagging

Manually tagged transcript	Tags (start tag, /end tag)													
	t	/t	s	/s	ss	/ss	mix	/mix	eng	/eng	j	/j	cd	/cd
M1	0	6	0	3	0	0	1	5	4	10	1	3	-	-
M2	0	2	0	0	0	0	0	2	528[a]	5	1	1	-	-
M3	0	0	0	0	0	0	1	0	0	0	0	0	-	-
M4	1	1	0	3	0	5	0	0	1	7	1	4	0	37
M5	0	1	0	0	0	14	1	10	0	5	1	5	-	-
Sum	1	10	0	6	0	19	3	17	633	27	4	13	0	37

Note: t = teacher; s = student; ss = students; mix = mixture of English and Japanese in utterances; eng = English; j = Japanese; cd = audio CD; M = Manually-tagged transcript. a. The participant that coded M2 misunderstood the tagging rule, and misplaced </eng> for start tags.

The next step (Step 4; Figure 4) used XSLT to extract the teacher's English utterances. The XPath to reach the teacher's English utterances; <xsl:copy-of select="body/t/eng"></xsl:copy-of>, extracted the utterances. Table 2 shows the extraction summary.

Table 2. XSLT summary: the number of lines of the teacher's English utterances

Transcript ID	Tagged by		Discrepancy (CCT-Manual)
	Manual	CCT	
1	266	202	-64
2	342	340	-2
3	73	62	-11
4	350	271	-79
5	144	117	-27

Note: XSLT = XSL transformation. CCT = the classroom corpus tagger.

As Ohashi, Katagiri and Oshikiri (2021) implied, the results revealed that more utterances could be retrieved through manual tagging than through CCT tagging. However, the manual tagging did not necessarily encompass the CCT tagging. The XSLT results from either one of the two tagging methods contained lines that did not appear in the other tagging method. Such lines were complementarily distributed in the XSLT results. Table 3 shows the summary of the complementary distribution quantity of the two tagging types.

Table 3. Summary of the complimentary line quantity extracted by XSLT

Transcript ID	Tagged by		Discrepancy (Manual + CCT)
	Manual	CCT	
1	73	7	80
2	6	1	7
3	13	1	14
4	79	0	79
5	31	3	34

The complementary lines, those not appearing in the counterpart tagging method, turned out to be of three types. Table 4 shows samples based on discrepancy types.

Table 4. XSLT extraction discrepancy due to tagging error types

Type	Tagging discrepancy sample
1 (Double-byte non-literal English characters)	T: It's Wednesday. CCT: <t><mix><eng>It</eng><j>'</j><eng>s Wednesday.</eng></mix></t> Manual:<t><eng>Yes, it's Wednesday. </eng></t>
2 (Single-byte characters for transcribing Japanese proper nouns)	T: He is XXX YYY. CCT: <t><eng>He is XXX YYY.</eng></t> Manual: <t><mix><eng>He is </eng><j> XXX YYY. </j></mix></t>
3 (Mis-tagging)	T: １グループで 1 eraser [Translation: One group can have one eraser.] CCT: <t><mix><j> １グループで </j><eng>1 eraser</eng></mix></t> Manual: <eng> １グループで 1 eraser</eng>

Note: XXX = a Japanese first name. YYY = a Japanese family name.

Type 1 tagging discrepancy resulted from the use of double-byte characters for apostrophes and the use of "smart/curly" double quotation marks. Type 2 resulted from different treatments of Japanese proper nouns, such as a person's name, being transcribed as English utterances. The transcriber took Japanese proper nouns to be Japanese, and thus, manually tagged them with <j> and </j>. However, the CCT tagged the Japanese proper nouns with <eng> and </eng>, as the CCT recognized the names represented by single-byte characters. The final discrepancy type, Type 3, resulted from tagging errors caused by manual tagging (Table 4). The error tagging example contains

a Japanese utterance, 1 グループ, indicating one group, with an English utterance, "easier", adding to the manual English tagging count. The manual coder should have initiated the tagging with <mix> followed by <eng>.

The three error types (Table 4) resulted from the use of double-byte character (Type 1), orthographical representation of L1 proper names (Type 2), and by missing tags, mostly end tags (Type 3). However, the discrepancies were narrowed down when we considered (1) the Type 1 error mixed utterances to be English utterances, as such mixed utterances contained nothing but English, and (2) the Type 3 error English utterances to be mixed utterances, (Table 5).

Table 5. Modified XSLT summary: the number of lines of the teacher's English utterances

Transcript ID	Tagged by		Discrepancy (CCT – Manual)
	Manual (initial count – Type 3 error English count)	CCT (initial count + Type 1 error mix count)	
1	263 (266-3)	271 (202+69)	8
2	342 (342-0)	342 (340+2)	0
3	73 (73-0)	71 (62+9)	-2
4	347 (350-3)	346 (271+75)	-1
5	142 (144-2)	146 (117+29)	4

Note: XSLT = XSL transformation. CCT = the classroom corpus tagger.

The errors were attributed to the CCT's tagging, not manual tagging. As for Type 2 errors, an explicit rule for encoding Chinese, Japanese, Korean (CJK) proper nouns can prove helpful in avoiding errors. For example, the CCT users can avoid tagging errors by consistently using Unicode CJK characters in encoding CJK proper nouns, instead of encoding them using single-byte alphabetical characters. Thus, tagging of transcripts by use of CCT can yield classroom corpora with reliable speaker and language tagging.

4.2 Participants' Reflection Survey

We procured information regarding the participants' reflections on the technical issues of CCT and their views on the potential for reflective practice in the development of teacher talk in the classroom. In the section below, we first address the participants' thoughts on the technical issues of CCT, after which we describe the CCT's potential to

train preservice English language teachers.

4.2.1 Advantages of the CCT

The survey first examined the CCT's technical issues. The participants completed a questionnaire on their impressions while engaging in the two tasks; manual tagging and tagging by CCT. The responses revealed that, according to the participants, the CCT is more advantageous than manual tagging in terms of tagging time, the cumbersomeness of nesting XML tagging structure, and the consistency of XML format. The participants reflected that manual tagging took a long time and was quite a troublesome task (participants 1, 2, and 4). Some participants also experienced difficulty locating exactly where to embed the <mix> tags in the nested structure within a single speaker turn (participants 1 and 5).

All the participants testified that tagging through CCT was far quicker than manual tagging, especially the language tags. However, they were bothered when they needed to select the speaker tags. This was partly because the speaker default values were "st," "sts," and "hrt," which forced them to add their own original speaker tags. The added tags involved more clicking to select the intended speaker tags. The individuals doing the tagging were expected to place the appropriate speaker tags in such cases (participants 4 and 5). Besides manual selection, CCT reduced the overall workload.

4.2.2 CCT's Potential for Reflective Practice

The survey also investigated whether the participants could utilize the CCT for their reflective practice. This sub-section discusses the potential use of reflection on the tagged transcript data, although reviewing the video-recorded data can contribute to reflective practice as well. The participants re-evaluated their English lessons by revisiting them again while tagging the classroom utterances. The questionnaire focused on the preservice teachers' reflections on the quantitative and qualitative use of the target language (English), instead of their first language (Japanese). The questionnaire also allowed the participants to express their views on any of the aspects which could not be covered by the survey.

By observing the use of target languages (L1 and L2), on the one hand, four out of five participants reported that they resorted to their first language, Japanese (L1),

rather than the target language, English (L2; participants 1, 2, 3, and 5). Some realized that they preferred speaking L1 when they supplemented the explanation in L2 (participants 1 and 5). On the other hand, one remarked that her students had more difficulty in understanding L2 grammar explanations (participant 5). This participant reflected that she had to switch codes between L2 and L1 by judging the students' reactions.

By observing the CCT tagging, the participants obtained insights into the quantity and quality of their L2 utterances. The majority of participants reflected that they could have spoken more while speaking in L2 (participants 1, 2, 3, and 5). One remarked that the preservice teacher would have spoken less, thus allowing the students more opportunities to speak L2 (participant 4). Another participant suggested that the preservice teacher refined the quality of classroom talk, despite the adequate L2 exposure toward the students during the lesson (participant 1). For example, the teacher would have used easier words and expressions besides being more concise (i.e., used terms less frequently). Another participant commented that the preservice teachers should be more careful while using articles and make a proper distinction between singular and plural nouns when they refer to nouns (participant 2). Another participant confirmed that the teacher's L2 pronunciation and grammar were correct (participant 5).

With respect to the participants' overall impressions of CCT after tagging their classroom transcripts, they were able to improve their understanding of the teacher talk, regardless of the use of language (L1, L2, and L1–L2 mixed language use). The elaborate and explanatory language of the preservice teachers when explaining L2 grammar and language tasks was the first to be noted. Two participants commented that preservice teachers unnecessarily explained the lesson points, which caused unnecessary confusion and resulted in a reduction of the activity time (participants 1 and 5). The participants benefited from observing their lesson transcripts that came out with the tagging. The tagged transcripts helped them understand the patterns of teacher talk, pronunciation, and grammar mistakes made while using L2. One participant mentioned that the understanding would lead to improvement in the classroom teacher talk (participant 5).

5. Discussion and Conclusion

This section summarizes the paper by illustrating the CCT's advantages, its utility in compiling classroom corpora, and its potential application in teachers' reflective practice.

5.1 CCT Advantages and Limitations

The CCT can be used as a tool for compiling classroom corpora that can contribute to the improvement of language classes through reflective practices. The CCT tagging replaces the time-consuming manual tag-attaching task, which facilitates researchers' and teachers' endeavors to compile classroom corpora.

This study highlights the three ways in which the CCT can assist potential users. These significant benefits of using CCT were highlighted by this study:

1) It performs time-consuming tag-attaching tasks with more accuracy in less time.
2) It prevents errors that are likely to occur as a result of manual tagging.
3) It appends more speaker tags than can be attached to the utterances, based on the class environment.

Therefore, the CCT is advantageous as it not only lessens the labor but also the time required for the compilation of classroom corpora once the CCT is supplied with the classroom transcripts. The CCT can assist in encouraging researchers and teachers to increase the use of classroom corpora, which will eventually facilitate teachers' reflective practice, besides providing more opportunities to corpus researchers.

Extracting specific utterances with tags attached by CCT, such as teachers' English utterances extracted by <teacher><eng>, or students' English utterances tagged as <student><eng>, CCT enables us to count the tokens of both teachers' and students' Japanese/English utterances. Tagged classroom corpora created by CCT provide the following information as the source of reflection:

1) Linguistic phenomena such as type and the extent of the language used in class.

2) The ratio of both teachers' and students' talk.

Teachers can use quantified utterances such as types and tokens gained from corpora to compare linguistic phenomena, including the diverse vocabulary usage levels, among different classes and examine the effectiveness of their input.

We are presently in the initial stage of developing CCT, and are aware that we need to expand the current tag set further, to encompass a higher hierarchical classroom discourse structure. As mentioned above, the participants stated that the relatively rich information tagged text allows for qualitative analyses in addition to quantitative comparison among classes. For example, qualitative observation of tagged utterances can examine the type or amount of teacher's talk that facilitates students' English production. Additional tags categorizing different input types are also likely to help investigate the effect of each input type that contributes to students' participation. These are some of the aspects that differentiate the CCT data from video-recorded materials.

As a rule, the CCT assigns <j> to double-byte characters and <eng> to single-byte characters. Therefore, <j> is displayed even if you enter other double-byte character languages, such as Chinese and Korean. Similarly, as long as it is a single-byte character, such as French and German, the tagging result will be displayed as <eng>. Possible tagging errors that occurred could have been avoided by controlling the transcribing method by a text editor. CCT users must note that the CCT is a software program exclusively designed to tag Japanese and English characters. This means that the CCT still requires improvements; which is a limitation of this study that should be addressed in future work.

Further improvements should consider following the text encoding initiative[4)] coding guidelines for speech transcriptions. Multiple attributes in speaker tags should be used to describe utterances' of speakers, and their use of language in tagging. One example might be to encode <sp id="1" who="teacher_1" type ="complete" lang="eng">Hello.</sp> instead of <t><eng>Hello.</eng></t> to specify a speaker, and the language use in one single tag.

Another limitation, but not the last, is that error counting the manual tagging may be called into question. In the pilot study, the five participants individually tagged five different transcripts. We could have obtained a more precise tagging error count caused by human tagging if the five participants had tagged the other four transcripts instead

of just one.

5.2 Application of Reflective Practice

As suggested by existing reflective practice literature, the reflective practice of the participants fostered their critical awareness of the need to improve effectiveness and aided in the development of their classes. The responses to our survey revealed its significance in critical reflection, which was evident in the positive comments. Participants confirmed that engaging in reflective practice through corpus compilation guided the teachers to evaluate themselves and identify their weak points. These findings conform to the findings of previous studies with regard to the benefits of the reflective cycle, in which educators observe, analyze, and develop action plans (e.g., Zwozdiak-Myers, 2012; Mann & Walsh, 2017).

Transcribed and quantified data, gained through the compilation of a classroom corpus, can prove to be a useful medium for prompting reflections for language teachers. Walsh (2013) mentioned that improving classroom interactions through reflective practice can prove to be an effective means for professional development. Thus, the CCT can assist teachers, especially preservice teachers, in corpus building that can provide valuable evidence for their reflective practice, eventually facilitating the effective development of their classes.

Mann and Walsh (2017) also pointed out that reflective practices require appropriate tools for collecting evidence to reflect upon. The CCT can serve as one of those tools, because it enables teachers to compile classroom corpora more easily, compared to other tools, and eventually helps to quantify utterances made by the teachers and students, thereby enriching reflective practice through observation of quantified classroom data. The results of the survey discussed in the preceding section substantiated this. The participants remarked that they were able to reflect on their language use with regard to the L1–L2 ratio, and L2 linguistic errors, such as grammar and pronunciation identified in teacher talks. The CCT provides easy access to numerical values of the transcribed text. The numerical information presents a clue that helps in a better choice of language or instruction in class. Furthermore, such data can supplement video-recorded materials in reflective practice.

This study demonstrates the advantages of CCT such as reduced painstaking manual tagging time required for transcription and also reduced errors associated with

the airtight start-end XML tag nesting structure. The classroom corpora complied in this manner could potentially provide the opportunity to the teachers for reflective practices.

CCT usage for the compilation of classroom corpora creates a pedagogical sequence of initial teaching, followed by reflection (reflective practice), and teaching again after reflection, when improvement will be expected. This primarily satisfies the need for classroom-related research, and also facilitates the professional development of language teachers. For example, quantified classroom data gathered through corpora quickly shows the number of vocabulary items included in tagged utterances; this can be used by researchers to compare different classes, enabling teachers to reflect on the improvement required in their teaching.

Expanding the use of CCT tagging, for example, to include the interactions and different tasks, will allow teachers to contemplate their classroom interactions more deeply and provide a wider opportunity for deep reflection. Such teachers' reflection through quantification of their utterances in a classroom will increase their awareness of what is needed to improve their classes. It will not only facilitate scaffolding but will also provide a more refined language exposure to their future students.

Acknowledgements

This study is part of a research program financed by Grant-in-Aid for Scientific Research (C) No. 19K00924. The authors cordially appreciate the insightful comments offered by anonymous reviewers in publishing this manuscript, which is based on the presentation and Proceedings of the JAECS 47th Conference 2021.

Notes

1. The CEFR-J Wordlist Version 1.6. Compiled by Yukio Tono, Tokyo University of Foreign Studies. Retrieved from http://www.cefr-j.org/download.html#cefrj_wordlist on 07/03/2021.
2. CCT download is available at: https://drive.google.com/file/d/1uD0MAwhiub1miKt1n-_qSw-Mmzira9Ek/view?usp=sharing
3. We used EditiX XML Editor 2015 for macOS.
4. Text Encoding Initiative guidelines (2014), 8.3 Elements Unique to Spoken Texts, *P5: Guidelines for Electronic Text Encoding and Interchange*, available at https://

tei-c.org/release/doc/tei-p5-doc/en/html/TS.html

References

Alger, C. (2006). 'What went well, what didn't go so well': growth of reflection in preservice teachers. *Reflective Practice, 7*(3), 287-301. https://doi.org/10.1080/14623940600837327

Anthony, L. (2015). TagAnt (Version.1.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved January 28, 2021, from www.laurenceanthony.net/software/tagant/

Bailin, S., Case, R., Coombs, J., & Daniels, L. (1999). Conceptualizing critical thinking. *Journal of Curriculum Studies, 31*, 285-302. https://doi.org/10.1080/002202799183133

Biber, D. (2010). Biber Tagger [Computer Software]. Flagstaff, AZ: Northern Arizona University.

Boud, D., & Walker, D. (1998). Promoting reflection in professional courses: the challenge of context. *Studies in Higher Education, 23*(2), 191-206. https://doi.org/10.1080/03075079812331380384

Harkin, J. (2005). Fragments stored against my ruin: the place of educational theory in the professional development of teachers in further education. *Journal of Vocational Education and Training, 57*(2), 165-179. https://doi.org/10.1080/13636820500200281

Katagiri, N., & Kawai, G. (2016). Designing XML schema for classroom discourse visual representation through XSLT. *Journal of Hokkaido University of Education (Humanities and Social Sciences), 66*(2), 1-16.

Katagiri, N., & Ohashi, Y. (2017). Analyses of Non-Native Preservice English Teacher Verbal Interactions at Japanese Middle Schools. *International Journal of Language Learning and Applied Linguistics World 15*(4). 1-16.

Katagiri, N., & Ohashi, Y. (2018). Developing Spoken Corpora of Non-Native English Teachers to Assist in English Classroom Interactions. *Official Conference Proceedings of the IAFOR International Conference on Language Learning*, 45-62.

Mann, S., & Walsh, S. (2017). *Reflective Practice in English Language Teaching*. Routledge.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies*. Routledge.

Ohashi, Y. (2015). A Corpus-based study on the relationship between the languages used in junior high school classrooms and learners' uptake. *KATE Journal, 29*, 29-42.

Ohashi, Y., & Katagiri, N. (2016). The effects of explicit instructions observed in teacher transcripts and student impression remarks in elementary school. *HELES Journal, 16*, 3-18.

Ohashi, Y., Katagiri, N., & Oshikiri, T. (2021). Developing Classroom Corpus Tagger: A Spoken Language Tagger to Compile Classroom Corpora. *Proceedings of the JAECS 47th Conference* 2021, 43-48.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From Corpus to Classroom*. Cambridge University Press.

Pollard, A. (2005). *Reflective Teaching*. London: Continuum.

Sinclair, J. M., & Coulthard, M. (1975). *Towards an Analysis of Discourse: The English Used by Teachers and Pupils.* Oxford: Oxford University Press.

Text Encoding Initiative. (2014). 8.3 Elements Unique to Spoken Texts, *P5: Guidelines for Electronic Text Encoding and Interchange*. Retrieved from https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Walsh, S. (2013). *Classroom Discourse and Teacher Development.* Edinburgh University Press.

Zeichner, K. M., & Liston, D. P. (1996). *Reflective Teaching: An Introduction.* Mahwah, N.J: Lawrence Erlbaum Associates.

Zwozdiak-Myers, P. (2012). *The Teacher's Reflective Practice Handbook*. *Becoming an Extended Professional Through Capturing Evidence-informed Practice.* London and New York: Routledge.

（大橋由紀子　ヤマザキ動物看護大学　Email: y_watanabe@yamazaki.ac.jp）
（片桐　徳昭　北海道教育大学　Email: katagiri.noriaki@a.hokkyodai.ac.jp）
（押切　孝雄　戸板女子短期大学　Email: oshikiri@toita.ac.jp）

「研究ノート」
『パラレルリンク』（Ver.1.0）の開発
ーパラレルコーパス研究の概観とコーパス整備ー

仁科　恭徳・赤瀬川史朗

Abstract

In this paper, we first review previous parallel corpora and analysis studies. We also suggest some future directions in this field. Then, we outline nine parallel corpora included in Parallel Link (Ver.1.0), an online analysis tool for Japanese-English/English-Japanese parallel corpora under development. In particular, we elucidated the text processing, annotation, creation of full-text search indexes, and file organization applied to these parallel corpora.

１．はじめに

本稿では，まず，現在までに構築された日英・英日パラレルコーパス，開発されたコンコーダンサー等の検索ツール，日英・英日パラレルコーパスを活用した研究を概観し，各パラレルコーパスの翻訳方向やテクストジャンル等の特徴と問題点, および今後ニーズが高まるツールや研究について具体的に述べる。次に，これからのパラレルコーパス研究の方向性を示すべく開発中の日英・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）に搭載予定の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，ファイル整理について詳説する。

２．パラレルコーパス研究の概観

2.1. パラレルコーパス開発研究

現在までに様々な日英・英日パラレルコーパスが開発されている。言語研究で使われてきた古い日英パラレルコーパスの一つに，関西外大コーパスＢ－

日英パラレルコーパス（西村，2002）がある。OS は Windows のみ，専用の検索プログラム Parallel Scan のみで検索可能であったが，2021 年 8 月時点で使用不可である。このコーパスでは，アライメント（例えば，英文 I'm Ken. / 和文「私はケンです。」を，各テクストファイルの同一行番号に配置する処理）が文単位ではなく段落単位となっているため，ParaConc（Barlow, 2002）などで読み込んで使うことはできない。他にも，言語（教育）研究で使われてきた代表的なものに日英新聞記事対応付けデータ（JENAAD）（Utiyama & Isahara, 2003）がある。読売新聞と Daily Yomiuri の対訳データ（一対一対応の日英文，一対多もしくは多対一対応の日英文）で，無料配布は既に終了している[1]。JENAAD を用いた研究には，日・英語間における交換可能性を量的に調査した仁科（2008a）や，英語教育的活用を目的として開発された LWP for ParaNews（公開終了）の授業利用に関する中條他（2015）がある（LWP については次節を参照）。表 1 は現在までに構築された日英・英日パラレルコーパスの例であり[2]，太字の 9 種のパラレルコーパスは，第 3 節で紹介する『パラレルリンク』（Ver.1.0）に搭載予定のコーパスを示す。表 2 はこれら 9 種の対訳対数と語数を示す。

表 1. 2000 年以降に公開された日英・英日パラレルコーパスのまとめ 時系列順（2021 年 9 月現在）

パラレルコーパス	先行研究	翻訳方向	アラインメント単位	ジャンル / レジスター	S/W
Tatoeba 日英対訳コーパス / 田中コーパス（TATOEBA）	Tanaka（2001）, Tatoeba project since 2006	日→英	文	教科書（例：日本人英語学習者が使用している書籍）／歌詞／一般書／聖書の一節	S・W
関西外大コーパス B 日英パラレルコーパス	西村（2002）	日→英	段落	文学／青空文庫と『新潮文庫の 100 冊』に収録されている作品，およびその英訳	W
日英対訳文対応付けデータ（TAIYAKU）	Utiyama & Takahashi（2003）	英→日 日→英（一部）	文	文学 ／ Project Gutenberg, 青空文庫，プロジェクト杉田玄白から 160 作品	W
ロイター日英記事の対応付け（REUTERS）	Utiyama & Isahara（2003）	英→日	文	新聞・ニュース／ロイター通信英・日本語版記事	W
JENAAD 日英新聞記事対応付けデータ	Utiyama & Isahara（2003）	日→英	文	新聞・ニュース／読売新聞，Daily Yomiuri	W
大規模オープンソース日英対訳コーパス（OPENSOURCE）	石坂他（2009）	英→日	文	技術文書／オープンソースソフトウェアのマニュアル	W
Wikipedia 日英京都関連文書対訳コーパス（WIKIPEDIA）	NICT（2010）	日→英	文	ウェブ／ Wikipedia の日本語記事（京都関連）とその英訳	W
TED Talk 日英コーパス（TED）	Cettolo, Girardi & Federico (2012) や https://wit3.fbk.eu/ で初期データ公開	英→日	文単位ではない ※字幕ファイル（ssa 形式ファイル）から作成	アカデミック・ビジネスプレゼンテーション／多種多様なジャンルのプレゼンテーションの字幕データ（音声ファイル付き）	S

日英法令対訳コーパス (LAW)	Neubig（2014）	日→英	段落	法律文書／日本の法律とその英訳	W
SCoRE 用例コーパス（SCoRE）	Chujo, Oghigian & Akasegawa（2015）	英→日	文	教育用例文／教育的に配慮した簡潔で自然な例文（音声ファイル付き）	W
ASPEC（Asian Scientific Paper Excerpt Corpus）	Nakazawa *et al.*（2016）	日→英	文	学術／科学論文の日英アブストラクト	W
Hiragana Times 日英対訳コーパスデータ	2017 年までのマガジン，別冊書籍の日英対訳データ YAC（Your Additional Contact）（https://yac-nippon.com/corpus-english-japanese/en/）	日→英	文（マガジン自体の表示形式は段落単位）	バイリンガルマガジン・書籍／1988 年から 2017 年までの Hiragana Times 349 冊，単行本 19 冊（政治，文化，歴史，恋愛，食べ物，旅行，映画，まんが，習慣など）	W
JESC（Japanese-English Subtitle Corpus）日英サブタイトルコーパス（の一部）	Pryzant *et al.*（2018）	英→日 日→英（一部）	文	映画・ドラマ・テレビ／映画・TV 番組の字幕データ	S

* S/W は Spoken/Written の略である。

表 2. 既存パラレルコーパスの対訳対数と語数（仁科・赤瀬川（2021）を参考。時系列順に改変）

コーパス名	対訳対	語数（日本語）	語数（英語）
TATOEBA	208,013	2,080,831	1,601,860
TAIYAKU	110,909	1,905,586	1,399,650
REUTERS	70,120	2,068,681	1,740,428
OPENSOURCE	505,780	6,927,281	5,018,603
WIKIPEDIA	443,849	9,132,894	9,806,199
TED	518,233	4,657,169	3,247,654
LAW	262,448	9,264,891	9,508,555
SCoRE	10,459	160,337	101,562
JESC	330,102	2,736,837	2,222,329
合計	2,459,913	38,934,507	34,646,840

表 1 から，まず，話し言葉のパラレルコーパスが少なく，最近になって構築され始めたことが分かる。また，この 10 年間でパラレルコーパスはいくつか開発されているものの，自然言語処理・機械翻訳分野で好まれる学術・技術系の専門的なコーパスが多いことも分かる。そして，一部のパラレルコーパスでは翻訳方向が考慮されず，日→英，英→日の双方向翻訳のテクストが混在していることや，アライメントが文単位のものと段落単位のものとに分かれていること，ジャンルに偏りが見られることなども指摘できる。無論，各々に利点があってのテクスト整形ではあるが，これら全てのコーパスを一定ルールのもと統一し合算できれば，ある程度のサイズが保証された擬似的な一般参照日英・英日パラレルコーパスとして言語分析等に活かすことができる[3]。

なお，JENAAD に関しては，2013 年から LWP for ParaNews が公開されオンライン上で簡易検索が可能となっていたが，2021 年 8 月時点で既に公開が終了している（LWP については後節を参照）。また，染谷・赤瀬川・山岡（2011）で使用された Wikipedia 日英京都関連文書対訳コーパス（https://alaginrc.nict.go.jp/WikiCorpus/）も日英パラレルコーパスがレキシカルプロファイラーに実装されている数少ないオンライン検索ツールで，NICT（情報通信研究機構）が 2010 年 10 月に一般公開した Wikipedia の日本語記事（京都関連）とその英訳から構成される日英パラレルコーパスであったが，2011 年 1 月の時点で開発が終了し（最新版は Ver.2.0.1），2021 年 8 月の時点で一般公開はされていない。京都に関する内容が中心で日本の伝統文化，宗教，歴史などの分野をカバーしている。人手翻訳による約 50 万文対を収録し（日本語の語数は約 1,000 万語），翻訳の過程（一次翻訳→流暢さ改善のための二次翻訳→専門用語チェックの 3

段階）を記録している。

表 1 に挙げた以外の日英・英日パラレルコーパスもいくつか存在しており，2021 年 8 月現在リンク切れもあるが日本語対訳データリスト（http://www.phontron.com/japanese-translation-data.php?lang=ja）が参考になるので参照されたい。また，Chujo, Oghigian, & Akasegawa（2015）や Mizumoto & Chujo（2016）など英語教育利用目的のパラレルコーパス検索ツール SCoRE の用例（http://www.score-corpus.org/download/jp/）もダウンロードすることができる。こちらは文単位でアライメントされており，計 10,459 件の英日対応文の用例を獲得することができる。ただし，教育利用が目的である当該コーパスは英文のレベルが統制されている点には留意されたい。

しかしながら，単言語コーパスと比較して日英・英日パラレルコーパスの構築には手間や時間を要することから，現状はその数と種類が限られている。表 1 から現状を把握すると，問題点として，翻訳方向，ジャンル構成比，検索ツール開発の 3 点が挙げられる。特に，今後は翻訳方向（日→英，英→日）ごとに，構成比やバランスも考慮しながら欠落しているジャンルやレジスターのコーパスを追加・構築する必要がある[4]。これは，各特定領域で使用される翻訳ユニットの抽出に有効であるだけでなく，ビジネス文書作成時における実務的利用や，様々な分野における二言語 DDL を用いた教育にも利用できる。そして，このような複数のパラレルコーパスをバランスよく合算することで，擬似的な一般参照パラレルコーパスとしての利用も可能となる。この場合，複数のパラレルコーパスを網羅的に串刺し検索できる使い勝手の良いツール開発も求められる[5]。仮にそのようなツールが開発できれば，二言語辞書に掲載すべき訳語や用例の信頼性と客観性が担保でき（仁科，2008a, 2020），辞書編纂時のコーパス活用の幅も広がる。

2.2. パラレルコーパス検索ツール研究

日英・英日パラレルコーパスの検索ツールには，コンコーダンサーとレキシカルプロファイラーがある。まず，現在使用可能なパラレルコーパス用コンコーダンサーに，Windows のみで使用可能な ParaConc（Barlow, 2002），Windows, Mac，Linux で使用可能な AntPConc（Anthony, 2017），Mac のみで使用可能な CasualPConc（Imao, 2018）がある。このうち，ParaConc のみ有償で，残りの二つは無料で使用できる。CasualPConc の姉妹版 CasualMultiPConc（最新版は Ver.0.4.1）では 2 ～ 5 言語の多言語コーパスの検索・処理が可能である

が，その開発は 2021 年 8 月時点で既に止まっている。前節で挙げた Parallel Scan（西村，2002）も現在使用不可である。

次に，パラレルコーパスが実装され，「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」（染谷・赤瀬川・山岡，2011）することが可能であるレキシカルプロファイラーに LWP（LagoWordProfiler）がある[6]。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（パルデシ・赤瀬川，2011）で，単言語コーパスを搭載した代表的なものに，現代日本語書き言葉均衡コーパス（Balanced Corpus of Contemporary Written Japanese: BCCWJ）の検索を可能とした NLB（NINJAL-LWP for BCCWJ）（パルデシ・赤瀬川，2011）がある。日英パラレルコーパスを搭載したものには，前述の JENAAD の検索が可能な LWP for ParaNews（2021 年 8 月時点で既に公開終了）や Wikipedia 日英京都関連文書対訳コーパスの検索が可能な Wikipedia-Kyoto LWP（WK-LWP）（2021 年 8 月時点で既に公開終了），そして SCoRE（オンライン公開中）がある。なお，Sketch Engine（https://www.sketchengine.eu/）もパラレルコーパスに対応しているが，現時点ではコンコーダンサーとしての機能のみである。

コンコーダンサーの使用には，検索の自由度が高い分，使用者の経験や分析力，直感力等が求められる。一方，レキシカルプロファイラーを使用した場合，あらかじめ決められた文法パターンにそった検索結果が表示されるため，敷居が低い分，検索や分析の柔軟性に欠けるという欠点もある。共に利点・欠点があることから，今後は双方の機能を実装したオンライン検索ツールを開発し無償公開することで一般ユーザーも含めて使用者の増加が見込めるのではないだろうか。特に，様々なジャンル・レジスターのパラレルコーパスをクリック一つで選択でき，検索項目の翻訳ユニットのジャンル構成比なども瞬時に表示できれば，今まで開拓されていなかった研究も可能となろう[7]。

また，京都外国語大学で展開している二言語同時学習（https://www.kufs.ac.jp/faculties/unv_education/unv_program_bi-language.html）をヒントに，開発が止まっている多言語対応コンコーダンサー CasualMultiPConc を用いることができれば，日本語・英語に加え，スペイン語や中国語といった 3 言語以上の複言語 DDL を外国語の授業で展開することもできる。英語に加えて他言語も学んでいる学習者にとって，マルチリンガルコーパスは活用に値するツールとなる。実際に多言語 DDL を教育現場で試した実践例はないため，その有効性を検証すべきであろう。単言語コーパスコンコーダンサーと比較してニー

ズは少ないが，CasualMultiPConc の開発を継続してもらいたい。

2.3. パラレルコーパス活用研究

仁科（2020）でもまとめたが，国内の日英・英日パラレルコーパスを活用した言語記述の研究は多くない。少し古くなるが，2002 年に発刊された『英語コーパス研究』第 9 号に掲載されている 9 本の論文はいずれも当時のパラレルコーパスの構築や検索プログラム，活用事例を知る上で極めて重要である。意味や文法，辞書学的見地から考察したものに，日英再帰形に注目した清水・村田（2002），身体部位を含む日英語表現を分析した岡田（2002），when 節を取り上げた田中（2002）などがある。2002 年以降では，その数は減り，時事英語表現の翻訳傾向などを調査した仁科（2006, 2008a, 2008b），カタカナ語の誤用を取り上げた Nishina（2008），日本語複合動詞とその翻訳を精査した染谷・赤瀬川・山岡（2011），依存木の統語構造的不一致から日英翻訳を分析した Oya（2017），そして，日本語動詞「固める」の翻訳ユニットを日本語コーパス（BCCWJ）と日英パラレルコーパス（WikipediaKyoto LWP）から分析した仁科（2020）などがある（自然言語処理や機械翻訳などの分野では研究報告が目立つ）。なお，これらの中でレキシカルプロファイラーを用いた研究は少なく，染谷・赤瀬川・山岡（2011）と仁科（2020）がそれにあたる。

一方，英語教育関係においては，ツール開発や DDL の教育的活用・効果検証等に関する研究，例えば WebParaNews（https://www.antlabsolutions.com/webparanews/about.html）の中條他（2014, 2015），Anthony, Chujo, & Oghigian（2011）や，SCoRE（http://www.score-corpus.org/）の Chujo, Oghigian, & Akasegawa（2015），Mizumoto & Chujo（2016）などの論文が発表されている。特に SCoRE に関しては，「慎重に作成した簡潔で自然な英語例文約 10,000 文と，日本人英語教師が丁寧に付けた日本語対訳文」から構成され，英文レベルが統制されている。日・英語の記述を分析する場合は，収録された英語テクストが制限・統制なしで産出されたものが理想であるため，言語研究と教育研究の利用とで求められるパラレルコーパスの質が異なることに留意されたい。

これからのパラレルコーパス研究に関しては，複数のパラレルコーパスが予め搭載され一括検索できる検索ツールの開発が進めば，日英・英日翻訳の量的分析が容易となり，特定の語・句のジャンルごとの翻訳実態の解明や，英和・和英辞書あるいは辞書データベースに掲載されている訳語・訳例を客観的に精

査することが可能となる。あるいは，一般的に我々が想定しているものとは異なる質のパラレルコーパスを用いた研究も考えられる。例えば，複数の翻訳家による翻訳ストラテジーの計量的調査を見込んだ一対多のパラレルコーパスの構築とその研究も興味深い。具体的には，同一の起点言語のテクスト（source text）と複数の翻訳家によって作成されたその目標言語のテクスト（target text）の各文をアライメントすることで構築される一対多のパラレルコーパスを活用すれば，各翻訳家の翻訳ストラテジーの類似性や相違性が可視化できる。熟達した翻訳家が暗黙に共有している翻訳ストラテジーの解明のみならず，各翻訳家の個性もデータとして可視化できるのである。参考までに，ルイス・キャロル著 *Alice's Adventures in Wonderland* には翻訳家 39 人による計 55 の日本語訳版が存在し，*Through the Looking-Glass* には 20 人の翻訳家による計 27 の日本語訳版が存在している（詳しくは，http://www.hp-alice.com/lcj/g_contents.html）。また，楠本（2001: 4-7）によれば，1998 年時点で両アリスの作品は 150 種前後が存在しているという説もある。これらの翻訳作品で用いられた表現や翻訳手法を計量的に比較するためには一対多のパラレルコーパスの構築が必要不可欠であり，筆者の知るところ，現時点でそのような研究は皆無である。今後の翻訳研究を前進させる上でその構築と分析には一定の価値があろう。

3．『パラレルリンク』（Ver.1.0）の開発準備

前節までの日英・英日パラレルコーパス（研究）の状況を受け，既存の日英・英日パラレルコーパスを串刺し検索できるオンライン検索ツール『パラレルリンク』（Ver.1.0）を Lago NLP（旧 Lago 言語研究所）と共同開発中である。仁科・赤瀬川（2021）が示すように，Ver.3.0 までを予定している約 10 年計画の本プロジェクトでは，最終的に各ジャンルにつき双方向翻訳のパラレルコーパスを搭載し，欠落しているジャンルのコーパスについては一からの構築も検討している。本節では，その第一段階として取り組んでいるプロトタイプに搭載予定のパラレルコーパスの選定と，それらのテクスト処理，アノテーション，全文検索インデックスの作成，ファイル整理について説明する。

3.1.『パラレルリンク』（Ver.1.0）に搭載予定のパラレルコーパス

本ツールの開発に先立ち，既存のパラレルコーパスの中身を再整備した。対象とした日英・英日パラレルコーパスは，表 1 中において太字で示した計 9 種

である[8]。大規模ウェブパラレルコーパス JParaCrawl（1,000 万対）（http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/）も収録を検討したが，ノイズが多いため今回は見送った。また，アカデミック分野のパラレルコーパス Asian Scientific Paper Excerpt Corpus（ASPEC）（300 万対）（http://orchid.kuee.kyoto-u.ac.jp/ASPEC/）の収録も検討したが，今回の『パラレルリンク』は一般公開を目指していることから，研究利用に限定されている当該コーパスは含めていない[9]。一方で，仁科（2020）では映画やドラマの日英・英日字幕コーパスを含める有用性に触れたが，今回，ノイズが多いものの JESC は含める方針を採った。また，Tatoeba コーパスに収録されている英和対訳文の元になった田中コーパスは，学生が翻訳した対訳文を数年かけて収集した約 15 万対のコーパスであるが，会話文が全体の 40% を占めることから含めることにした[10]。

なお，教育目的では SCoRE 用例コーパスを活用することが最適であるが，言語学的な分析に関しては今回選定した 9 種の中では，SCoRE を除く他 8 種のパラレルコーパスを用いるべきかもしれない（仁科，2020 参照）。また，各コーパスサイズが異なることから，コーパス間で比較分析（ジャンル・レジスター分析）を可能にするために，コーパスごとに 100 万語あたりの生起頻度を表示する機能をインターフェースに実装する予定である。しかしながら，今回搭載した 9 種のコーパスだけでは検索したい語・句の十分な翻訳例が得られない可能性もあるため，表 1 に挙げた他のパラレルコーパスや自作コーパスの追加も Ver.2.0 以降に検討したい[11]。

3.2.『パラレルリンク』（Ver.1.0）のテクスト処理・アノテーション

次に，各パラレルコーパスのフォーマットを統一するためにテクスト処理を施し，英語には品詞情報，日本語には形態素情報を付与した。そして，Blacklab Query Tool（https://inl.github.io/BlackLab/query-tool.html）を用いて全文検索のインデックスを作成した。まず，テクスト処理としてテクストのクリーニング，エンコーディングの統一（UTF8），フォーマットの統一，センテンス ID の付与を施した。以下は，対訳ファイルのサンプル（TED）である。次に，品詞情報・形態論情報を付与した。まず，英文に関しては，Stanford POS Tagger（https://nlp.stanford.edu/software/tagger.shtml）を用いて，表層形，レマ，品詞など品詞に関する情報を付与した。また，日本語に関しては形態素解析器 Sudachi（https://github.com/WorksApplications/

Sudachi）を使用し，表層形，語彙素，品詞に関する形態論情報を付与した。今後，文単位への変換を予定している。

```
TED 00001 0000000001 I'm going to talk to you tonight 今晩 お話しするのは
TED 00001 0000000002 about coming out of the closet カミングアウトについてです
TED 00001 0000000003 and not in the traditional sense いわゆる「カミンアウト」
TED 00001 0000000004 not just the gay closet. ゲイだと打ち明けることではありません
TED 00001 0000000005 I think we all have closets. 誰しも心に壁を作っています
TED 00001 0000000006 Your closet may be telling someone その後ろに隠れているのは
TED 00001 0000000007 you love her for the first time 誰かに初めて愛の告白をすることや
TED 00001 0000000008 or telling someone that you're pregnant 妊娠したこと
TED 00001 0000000009 or telling someone you have cancer ガンであることを伝えることかもしれません
TED 00001 0000000010 or any of the other hard conversations 他にも私たちが人生で経験する一
```

図 1. 対訳ファイルサンプル（TED）

3.3.『パラレルリンク』（Ver.1.0）の全文検索インデックスの作成

その後，全文検索インデックスを作成した。詳しくは，上記の品詞情報，形態論情報を含む Blacklab Query Tool のインポートファイルを作成した。ファイル形式は XML である。

```
<?xml version="1.0" encoding="UTF-8"?>
<docs>
 <doc corpus="TED" subcorpus="" fid="00001" sid="0000000001" type="en" counterpart="
今晩 お話しするのは ">
  <s id="TED:;00001:0000000001">
   <w p="PRP" l="I">I</w>
   <w p="VBP" l="be">'m</w>
   <w p="VBG" l="go">going</w>
   <w p="TO" l="to">to</w>
   <w p="VB" l="talk">talk</w>
   <w p="TO" l="to">to</w>
   <w p="PRP" l="you">you</w>
   <w p="RB" l="tonight">tonight</w>
  </s>
 </doc>
```

図 2. インポートファイルサンプル（TED 英文）

```
<?xml version="1.0" encoding="UTF-8"?>
<docs>
  <doc corpus="TED" subcorpus="" fid="00001" sid="0000000001" type="ja"
counterpart="I&#x27;m going to talk to you tonight">
    <s id="TED:;00001;0000000001" corpus="TED" type="ja">
      <w p=" 名詞 , 副詞可能 ,*,*" l=" 今晩 "> 今晩 </w>
      <pu> </pu>
      <w p=" 名詞 , サ変接続 ,*,*" l=" お話し "> お話し </w>
      <w p=" 動詞 , 自立 ,*,*" l=" する "> する </w>
      <w p=" 名詞 , 非自立 , 一般 ,*" l=" の "> の </w>
      <w p=" 助詞 , 係助詞 ,*,*" l=" は "> は </w>
    </s>
  </doc>
```

図 3. インポートファイルサンプル（TED 日本語文）

3.4. ファイル整理

処理後のテクストファイルについては，大きく分けて 3 種のフォルダ（formatted; annotated; blacklab）に整理した。まず，formatted フォルダには，パラレルコーパスの種類ごとに 9 種類のサブフォルダ（JESC; Law; OpenSource; Reuters; SCoRE; Taiyaku; Tatoeba; TED; Wikipedia）が用意されている。また，各サブフォルダには，それぞれ以下 2 種類のテクストファイルが収録されている。コーパスデータは，コーパス，サブコーパス，ファイル ID，センテンス ID，英文，和文の 6 つのフィールドから構成され，各フィールドはタブで区切られている。エンコーディングは前述のとおり，UTF-8 で統一している。リンクデータは，コーパス，サブコーパス，ファイル ID，ファイル名の 4 つのフィールドから構成され，各フィールドはタブで区切られている。検索ツールを開発するときに，元のコーパスファイルを表示するために用いられる。また，original サブフォルダも用意し，こちらには変換前の元データが収録されている。

```
[ コーパス名 ].txt... 統一フォーマットのコーパスデータ
[ コーパス名 ].metadata.txt... 元のコーパスファイルとファイル ID とのリンクデータ
```

図 4. 各サブフォルダに収録されている 2 種類のテクストファイル

次に，annotated フォルダには，formatted フォルダにある統一フォーマットのコーパスデータにアノテーション情報を付与したファイルを収納している。ファイルのフォーマットは前述のとおり XML ファイルで，Blacklab Query Tool のインポートファイルとなる。各コーパスについて，英文と和文の 2 種類の XML ファイルが用意されている。

最後に blacklab フォルダには，Blacklab Query Tool のインデックスファイルが収録されている。検索ツールのバックエンドの役割を果たす。以上のようなテクスト処理，アノテーション，全文インデックス作成，ファイル整理を行うことで，既存パラレルコーパスを整備した。

4. まとめ

本稿では，はじめに過去から現在までの日英・英日パラレルコーパスや検索ツール，それらを活用した研究の変遷を振り返り，これからの展望を述べた。そして，現在開発中の日英・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）に搭載予定のパラレルコーパスの概要とテクスト処理などの一連の再整備作業について詳説した。当該検索ツールのインターフェースや実装している検索機能の紹介，活用研究などについては，論を改める。

謝辞

本研究は JSPS 科研費 20K00692 の助成を受けたものである。また，2021 年 10 月 2 日にオンラインにて開催された英語コーパス学会第 47 回大会において口頭発表したものに，大幅な加筆・修正を施したものである。ここに，『パラレルリンク』の開発に携わって頂いた第二著者の Lago NLP（旧 Lago 言語研究所）の赤瀬川史朗代表，および 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），ならびに関係者の皆様に感謝の意を示す。

注

1. 詳しくは，http://www2.nict.go.jp/astrec-att/member/mutiyama/index-ja.html を参照。

2. カーネギーメロン大学の Graham Neubig 氏のウェブサイト http://www.phontron.com/japanese-translation-data.php?lang=ja も参考にした。
3. 一般参照と呼ぶには，サブコーパスのサイズやバランス，構築デザインそのものを統制することが難しいため，「擬似的な」ということばをここでは用いた。
4. 翻訳物ということを考えると，収集テクストの時代性を統一することが難しく，原著のみならず翻訳物の版権の問題もあるため，そのハードルは想像以上に高い。
5. 既存のパラレルコーパスを再整備して一覧検索を可能にすることが，現時点でできる第一歩であろう。これが，第 3 節で紹介する『パラレルリンク』（Ver.1.0）の開発へと繋がっている。
6. 赤瀬川・パルデシ・今井（2014, p.41）によれば，レキシカルプロファイリングとは「あらかじめ設定された検索式に基づいて，コーパスから様々なタイプのコロケーションの情報を抽出した結果を，文法パターンごとに整理してユーザに提示するコーパス検索手法」であり，「特定の語彙の文法的振る舞いやコロケーションをマクロ的視点から調査できる」と説明する。
7. 第 3 節で紹介する『パラレルリンク』（Ver.1.0）では，手始めにレキシカルプロファイラーの機能を実装する予定であるが，Ver.2.0 以降ではコンコーダンサーも搭載する予定である。
8. コーパス・デザインから構築，版権取得までかなりのハードルがあるため，単言語コーパスと異なりパラレルコーパスの新たな開発や普及はそれほど進んでいない。よって，現時点で何ができるかを考えた場合，既存資源を有効活用する『パラレルリンク』には一定の意味があろう。
9. 内部利用はおそらく可能であることから，使用者を限定した研究者用のツール開発も進めたい。
10. 正確には 146,784 文が日本語と英語の両方で書かれており，大半が短文であり，英文の長さが平均で 7.72 語，最長で 45 語との報告がある（http://hihan.hatenablog.com/entry/2019/01/20/070254）。また，学生 1 人あたり 300 個の文章を翻訳したことから，翻訳者が多数存在する一方で複数の日本人大学生が翻訳プロジェクトに参加したため誤訳が混ざっている可能性もあり，質の点では保証できないという欠点がある。
11. ただし，JENAAD は既に無償配布が終了しているため，使用許諾に費用が発生する（JENAAD の有償ライセンスは非商用で 50 万程度である）。同様に Hiragana Times 日英対訳コーパスデータのアカデミックユースは一般の 40% 引きの 150 万程度で契約可能である。

参考文献

赤瀬川史朗・パルデシプラシャント・今井新悟（2014）「NINJAL-LWP の類義語比較機能」『第 6 回コーパス日本語学ワークショップ予稿集』41-50.

Anthony, L. (2017) AntPConc (Ver.1.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.html

Anthony, L., K. Chujo and K. Oghigian (2011) "A Novel, Web-based, Parallel Concordancer

for Use in the ESL/EFL Classroom." In Newman, J., H. Baayen and S. Rice (eds.), *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. New York: Rodopi, pp. 123-138.

Barlow, M. (2002) ParaConc: Concordance Software for Multilingual Parallel Corpora. [Computer Software]. Available from https://paraconc.com

Cettolo, M., C. Girardi and M. Federico. (2012). "WIT3: Web Inventory of Transcribed and Translated Talks." *Proceedings of the 16th EAMT Conference, 28-30 May 2012*: 261-268.

中條清美・アントニローレンス・内山将夫・西垣知佳子（2014）「フリーウェア WebParaNews オンライン・コンコーダンサーの英語授業における活用」『日本大学生産工学部研究報告 B』第 47 号：49-63.

中條清美・西垣知佳子・赤瀬川史朗・内山将夫（2015）「レキシカル・プロファイリング型オンラインコーパス検索ツール LWP for ParaNews の英語授業における利用」『日本大学生産工学部研究報告 B』第 48 号：45-57.

Chujo, K., K. Oghigian and S. Akasegawa (2015) "A Corpus and Grammatical Browsing System for Remedial EFL Learners." In Leńko-Szymańska, A., and A. Boulton (eds.), *Multiple Affordances of Language Corpora for Data-driven Learning*. Amsterdam: John Benjamins, pp. 109-128.

Imao, Y. (2018) CasualPConc (Ver.1.0) [Computer Software]. Available from https://sites.google.com/site/casualconcj/ その他のアプリケーション /casualpconc

石坂達也・内山将夫・隅田英一郎・山本和英（2009）「大規模オープンソース日英対訳コーパスの構築」『情報処理学会研究報告』第 17 号：1-7.

楠本君恵（2001）『翻訳の国の「アリス」－ルイス･キャロル翻訳史･翻訳論』未知谷 .

Mizumoto, A., and K. Chujo (2016) "Who is Data-driven Learning for? Challenging the Monolithic View of its Relationship with Learning Styles." *System 61*: 55-64.

Nakazawa, T., M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi and H. Isahara (2016) "ASPECT: Asian Scientific Paper Excerpt Corpus." *Proceedings of the 9th International Conference on Language Resources and Evaluation*: 2204-2208.

Neubig, G.（2014）日英法令対訳コーパス . Available from http://www.phontron.com/jaen-law/index-ja.html

NICT（2010）Wikipedia 日英京都関連文書対訳コーパス（Ver.2.0.1）. Available from https://alaginrc.nict.go.jp/WikiCorpus/

西村公正（2002）「誌上シンポジウム 日英パラレルコーパスでどのような英語研究が可能か -- コーパス構築の概要と検索プログラム , および研究事例」『英語コーパス研究』第 9 号：37-43.

仁科恭徳（2006）「相互関係を表す形容詞から見たシノニム学習の理論と実践教材 : 実証的考察とパラレルコーパスを用いたデータ駆動型学習法を中心に」『LET 関西支部研究集録』第 11 号：45-59.

仁科恭徳（2008a）「パラレルコーパスを用いた交換可能性の一考察」『英語コーパス研究』第 15 号：81-95.

仁科恭徳（2008b）「パラレルコーパスを用いた抽象語彙・フレーズの一考察 : これからの二言語辞書の編纂論」『LET 関西支部研究集録』第 12 号：83-97.

Nishina, Y. (2008) "Parallel Corpora in Computer-assisted Language Learning: A Case of Lexical Studies and Data-driven Learning Using Moodle". In Marriott, R., and P. Torres (eds.), *Handbook of Research on E-Learning Methodologies for Language Acquisition.* Hershey: Information Science, pp. 203-217.

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･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.

岡田啓（2002）「「顔」を含む日本語表現と対応する英語表現について」『英語コーパス研究』第 9 号：57-79.

Oya, M. (2017) "Syntactic Divergence Patterns among English Translations of Japanese One-word Sentences in a Parallel Corpus." *English Corpus Studies 24*: 19-40.

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』国立国語研究所，pp. 205-216.

Pryzant, R., Y. Chung, D. Jurafsky and D. Britz (2018) *JESC: Japanese-English Subtitle Corpus*. Ithaca, New York: Cornell University. Available from https://arxiv.org/pdf/1710.10639

清水眞・村田真樹（2002）「パラレルコーパスを用いた日英再帰形の分析」『英語コーパス研究』第 9 号：17-34.

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.

田中美和子（2002）「『語り』の when 節」の意味特徴」『英語コーパス研究』第 9 号：81-91.

Tanaka, Y. (2001) "Compilation of a Multilingual Parallel Corpus." *Proceedings of PACLING 2001*: 265-268.

Utiyama, M., and H. Isahara (2003) "Reliable Measures for Aligning Japanese-English News Articles and Sentences." *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics 2003*: 72-79.

Utiyama, M., and M. Takahashi (2003) *English-Japanese Translation Alignment Data*. Available from https://www2.nict.go.jp/astrec-att/member/mutiyama/align/index.html

（仁科　恭徳　神戸学院大学　Email: ynishina@gc.kobegakuin.ac.jp）

（赤瀬川史朗　Lago NLP（旧 Lago 言語研究所）Email: shiro.akasegawa@lagonlp.jp）

「研究ノート」

知覚動詞補文に出現する受身表現の容認可否について

村岡宗一郎

Abstract

Regarding the use of the non-finite verbs in the complement of perception verbs, “be + -en” is unacceptable, while “{being / get} + -en” are acceptable. However, unacceptable forms such as “see NP be + -en” are used in practice, for example, “I couldn’t stand to ***see her be cremated***. (Murakami Haruki, *Killing Commendatore*.)”. This study analyzes how often these example such as “see NP be + -en” are used in reality and what semantic constraints are imposed upon them by examining data from corpora, such as BNC and COCA. This study confirms that “see NP be + -en” is used more often in American than in British English, and “see NP get + -en”, which has been considered by previous studies to be grammatically correct, is also mainly used in American English. I conclude that the use of “see NP be + -en” has increased along with the use of “see NP get + -en” in American English.

1．はじめに

現代英語における知覚動詞は，（1a-b）に示すように，補文に原形不定詞，現在分詞，過去分詞をとる。このうち，原形不定詞は当該事象の全体性や完結性を表し，現在分詞は一時性や非完結性を表す（cf. Allen（1974[4]: 186））。過去分詞は当該動詞がもつプロセスの終点に焦点をあて（cf. Langacker（2008: 121）），受身を表すが，（1c）は一般的に容認されていない。

（1）a. I ***saw the children eat(ing)*** their lunch. （Palmer 1987[2]: 199）
b. I ***saw the children (being) beaten*** by their rivals. （ibid.）
c. *I ***saw him be rejected***. （Bolinger 1974: 69）

しかし，村上春樹の『騎士団長殺し』の英訳には，I couldn't stand to ***see her be cremated***. という例が確認されており，安藤（2005: 829, 2008: 113）は稀な例とする。本研究では（1c）の be -en 補文がなぜ容認されないのか，また be -en 補文が稀であるという言語事実に関して，どのような制約が課されるのかを明らかにしていく。本論考の構成は以下の通りである。まず，第 2 節では，知覚動詞補文における受身表現について先行研究の分析をまとめていく。そして，第 3 節では，先行研究の分析をもとに，BNC と COCA を用いて知覚動詞補文における受身表現の分布を調査し，第 4 節では，調査結果について考察していく。

2．知覚動詞補文における受身表現について

be -en 補文を容認する先行研究は（2）のように少数であり，Burzio（1986: 312）などの先行研究では（3）のように，一般的に非文法的であるとされてきた（cf. Bolinger（1974: 69）, Miller（2002: 249）, Basilico（2003: 9）, Dixon（2005[2]: 252））。

（2）a. I ***saw her be killed***. （Wilder 1992: 215）
b. I ***saw the dogs be all called*** back by their owners. （Guasti 1993: 133）
c. I {***saw / heard***} ***the teachers be fired***. （Sheehan and Cyrino 2018: 3）

（3）a. ?Mary ***saw the princess be kissed*** by the frog. （Lapointe 1980: 772）
b. We ***saw the dog (*be) run*** over by lorry. （Declerck 1991: 490）
c. *?John ***saw Bill be examined*** by a doctor. （Clark and Jäger 2000: 19）
d. *Jane ***saw Peter be kissed***. （Gisborne 2010: 209）

see 以外の知覚動詞について，Akmajian（1977）は（4a）に見られるように，watch NP be -en もまた非文法的であると分析するが，Lapointe（1980: 722）では認められている。hear の場合は，Declerck（1991）や Dixon（2005[2]: 252）は be の削除は義務的であるという。

（4）a. *We ***watched the rebels be executed*** by the army. （Akmajian 1977: 440）
b. I've never ***heard it (*be) said*** before. （Declerck 1991: 490）

その一方で，安藤（2005: 829, 2008: 113）は，Palmer（1965: 171）の例を引用し, be を付けるのはまれであると述べ容認しているが, 引用元の Palmer（1965: 171）の記述は Palmer（1974, 1987[2]）から削除されているため，この分析については検証する必要がある。このように，原形不定詞補文において be -en 補文はなぜ現代英語では非文法的とみなされる傾向にあるのだろうか。次節以降，be -en 補文の容認可否性について考察し，分析を加えていく。[1]

３．知覚動詞補文における受け身表現の容認可否

まず，知覚動詞の原形不定詞補文における be -en の出現について，（３）で見たように，従来非文法的であるとされてきた。Dixon（2005[2]: 252）によれば，be -en 補文の主語は知覚事象を開始する参与者ではないため容認されないと分析する。その一方で，Bolinger（1974: 69）や Miller（2002: 249）などの先行研究によれば（５）に示すように，受身を表す get -en は容認されている。また watch にも同様の容認可否性が見られる。

（５）a. John ***saw Bill*** {***get*** / ***?be***} ***examined*** by a doctor.　（Clark and Jäger 2000: 19）
b. We ***watched the rebels*** {***get*** / ***be***} ***executed*** by the army.（Akmajian 1977: 440）

これは状態的な be と動作的な get の違いであると考えられる。[2] このように状態受動が容認されない理由について，知覚動詞は（６）のように目に見える動作や状態のみを知覚対象とするため，状態性の強い be -en 補文は非文法的となる傾向にあると考えられる。

（６）Martha ***saw the policeman*** { ***nude / *intelligent / run(ning) into the bar / *own a car / *nice guys to old ladies / be(ing) heroes / chased by the Robbers / *be mammals / in the cruiser / with the monster / *liked by the robbers.*** （Carlson 1977: 125）

吉良（2006: 46）もまた，（７）のような原形不定詞補文における状態動詞の出現について，「状態的な出来事」は終結点を持つ「完了した事象」とは捉えられず，容認できないという。

（７）a. *We ***saw John look*** pretty sick. （Akmajian 1977: 440）
b. *I ***saw Tom*** still ***resemble*** your father. （Declerck 1981: 89）

柏野（1993: 80）は，原形不定詞を用いるということは，補文動詞を抽象化するということになると述べ，補文動詞の抽象化には一定の制限があり，通例，開始点と終結点として捉えられる動詞や瞬間動詞のように行為そのものが点として捉えられる動詞が用いられ，（8a）の look のように，開始点と終結点のはっきりしない状態動詞を原形不定詞で抽象化するのは，まず不可能だという。また（２）と（３）で見たように，be -en 補文において文法性の判断に差が見られる要因については，抽象化，つまり，何を点と見るかについては人により，あるいは文脈により差があるためであるという。また動作性を表す be -en 補文について，Bolinger（1974）は（８）のように習慣や反復を表す場合には，be -en 補文も容認されると述べるが，これは個々の状態的な出来事が繰り返しの動作として捉えられるためである。[3]

（８）a. I used to ***see him be rejected***. （Bolinger 1974: 69）
b. Again and again I ***saw him be rejected.*** （ibid.）

また知覚動詞補文においては，（６）で見たように，Martha ***saw the policeman be mammals***. の様な個別レベル述語は容認されないが，be が非状態性を表している場合には，原形不定詞補文における be -en の出現は容認されうる。このことについて，白井（1999: 20）は振る舞うというような動作的な be であれば（9a）は適格であると述べ，中右（1980: 148）も同様に，be -en が非状態的な事態を表している場合には，（9b）は適格文であると分析する。

（９）a. We ***saw John be polite*** for the first time. （Arimoto 1989: 119）
b. I don’t like to ***see people be intimidated***. （中右 1980: 147）

さらに，Bolinger（1974）や柏野（1993: 81）は，（10）のように知覚動詞の完了形であれば，原形不定詞補文においても，be -en の出現が容認されるという。

（10）I ***have seen him*** {***get / be***} ***rejected***. （Bolinger 1974: 69）

現在完了には「完了・結果」と「経験」の二つの解釈が可能であるが，柏野（1993: 78）によれば，主文が完了形で「完了・結果」の意味の場合には，主文に示される知覚過程の完結を強調するので，補文の行為も終わっていることを表すため，原形不定詞が選ばれるという。その一方で，補文の動詞が状態動詞で主文の動詞の時制が「経験」を表す完了形であれば，状態動詞（be を含む）も補文に生起可能となると述べ，主文が過去形の場合には，一般に抽象化のできなかった状態動詞も，主文が抽象化された概念を表す「経験」の完了形になると，完了形の意味そのものが抽象化された概念なので，その影響を受けて抽象化が可能になるという（cf. 柏野（1993: 79, 81））。[4] 以上の先行研究をまとめると，知覚動詞補文における be -en の出現は，be -en が動作的なものや知覚動詞の完了形が用いられている場合には容認される。次節以降，知覚動詞補文における be -en の使用について調査を行う。

4．調査概要と結果

前節にて，先行研究の見解をもとに知覚動詞補文における be -en の容認可否について分析し，容認される条件として知覚事象の動作性の強化という条件のもとで知覚動詞補文における be -en の出現が容認されうることが明らかになった。これらの先行研究の結果がどれほど実際の言語使用を捉えられているのかを検証するために，BNC と COCA を用いて調査を行う。[5] 調査を行うにあたって用いる検索式については，補文主語のバリエーションを考慮し，（11）のように，“（形容詞＋）名詞”，“代名詞”，“（不）定冠詞＋（形容詞＋）名詞” の 7 パターンに，複合名詞句を形成していると考えられる “（（不）定冠詞）＋（形容詞＋）名詞＋名詞” の 6 パターンを加えた計 13 パターンの名詞句を対象として検索した。[6]

（11）a. {[see] / [hear] / [watch]} (ART / DET) (ADJ) NOUN be _v?n
　　b. {[see] / [hear] / [watch]} PRON be _v?n
　　c. {[see] / [hear] / [watch]} (ART / DET) (ADJ) NOUN NOUN be _v?n

これらの検索式を用いて，知覚動詞 see の補文内部における受身表現の分布を調べた結果，イギリス英語では，be -en は少なく，being -en が圧倒的であった。また従来先行研究で容認されていた get -en の出現についてもイギリス英語に

は殆ど検出されなかった。その一方で，アメリカ英語では，get -en の用例が多く，be -en もまたイギリス英語より多く検出された。

表 1．BNC と COCA の see の補文における受身表現の分布

	BNC		COCA	
[see] NP be -en	**4**	**1.6%**	**118**	**4.5%**
[see] NP being -en	233	95.9%	1698	64.4%
[see] NP get -en	**3**	**1.2%**	**647**	**24.5%**
[see] NP getting -en	3	1.2%	174	6.6%
TOTAL	243	100.0%	2637	100.0%

see NP {be / get} -en の用例は（12）に示す通りである。COCA の be -en 補文には，主に born, made, used, prepared, taken が用いられ，一方で get -en 補文には主に hurt, hit, killed, arrested, shot, beaten が用いられていた。

（12）a. I want to ***see the baby be born***. （COCA: 2013. SPOK）
b. Did you ***see him get hit*** in his face? （COCA: 1999. MOV）
c. I would like to ***see the scheme be taken*** on, （BNC: H49. S_meeting）
d. they all thought theyd ***see us get thrashed***. （BNC: J1D W_email）

watch について同様の調査を行ったところ，表 2 に示す結果が得られた。イギリス英語においては，see の場合と同様に {be / get} -en の出現は殆ど見られなかった。しかし，アメリカ英語では，{be / get} -en の出現はほぼ同じ割合で検出された。これは，watch が動作的なものを補部にとるためであると考えられる。

表 2．BNC と COCA の watch の補文における受身表現の分布

	BNC		COCA	
[watch] NP be -en	**0**	**0.0%**	**199**	**21.0%**
[watch] NP being -en	71	97.3%	509	53.7%
[watch] NP get -en	**1**	**1.4%**	**220**	**23.2%**
[watch] NP getting -en	1	1.4%	19	2.0%
TOTAL	73	100.0%	947	100.0%

watch NP {be / get} -en の用例は（13）に示す通りである。COCA の watch NP be -en には，主に killed, destroyed, taken, murdered, born が用いられ，一方で watch NP get -en には主に beaten, slaughtered, blown, killed, hit, knocked が用いられていた。

（13）a. I won't stand by and ***watch him be executed***.　（COCA: 2019. TV）
b. We're not gon na sit around and ***watch them get slaughtered***.
（COCA: 2000. TV）
c. we ***watched surface responsibility get peeled*** away from a lot of people...
（BNC: K2R. W_newsp_other_arts）

hear は表 3 に示すように，{be / get} -en の例はイギリス英語では検出されず，アメリカ英語においてもほとんど検出されなかった。hear NP {be / get} -en の用例は（14）に示す通りである。COCA の be -en 補文には，主に called, compared, taken が用いられ，get -en 補文に特徴的なものは見られなかった。

表 3. BNC と COCA の hear の補文における受身表現の分布

	BNC		COCA	
[hear] NP be -en	**0**	**0.0%**	**18**	**7.8%**
[hear] NP being -en	50	98.0%	187	81.0%
[hear] NP get -en	**0**	**0.0%**	**12**	**5.2%**
[hear] NP getting -en	1	2.0%	14	6.1%
TOTAL	51	100.0%	231	100.0%

（14）a. I've ***heard you be compared*** to Nico.　（COCA: 2012. BLOG）
b. I've ***heard them get called*** some pretty mean names.　（COCA: 2012. BLOG）

これらの調査結果は，以下のようにまとめられ，be -en 補文と get -en 補文はどちらも英米に差が見られるという結果が得られた。

表 4. BNC と COCA の知覚動詞補文における受身表現の分布

	BNC		COCA	
be -en	**4**	**1.1%**	**335**	**8.8%**
being -en	354	96.5%	2394	62.8%
get -en	**4**	**1.1%**	**879**	**23.0%**
getting -en	5	1.4%	207	5.4%
TOTAL	367	100.0%	3815	100.0%

このように，be -en と get -en はアメリカ英語に多く見られるが，COCA で検出された用例の殆どが表 5 に示すように，被害を表す動詞の過去分詞が用いられていた。

表 5. COCA における “{see / watch / hear} NP {be / get} + -en” の過去分詞

{see / watch / hear} NP be + -en			{see / watch / hear} NP get + -en		
順位	件数	過去分詞	順位	件数	過去分詞
1	16	born	1	119	**hurt**
2	13	**destroyed**	2	55	**hit**
3	14	**killed**	3	35	**killed**
4	12	taken	4	23	**beaten**
5	7	**murdered**	5	19	**arrested**

また be -en 補文に関して，（10）で見たように先行研究では完了形で用いられる場合には容認されうるとされていたが，表 6 に示すように，COCA では be -en 補文と get -en 補文ともに約 1 割程度しか検出されなかった。その一方で，（15）に見られるように (don’t) like や hate などの好き嫌いを表す語の補部として用いられている例や感情表現と共起する例が多く検出された。さらに by によって動作主が具現化された例も多くは確認されなかった。

表 6.　{be / get} -en 補文と完了形・感情表現・by + 動作主の共起とその割合

	BNC			COCA		
総数（英 / 米）	完了形	感情表現	by	完了形	**感情表現**	by
be -en（4/335）	0	3（75%）	0	34（10%）	**73（22%）**	27（8%）
get -en（4/879）	0	0	0	90（10%）	**270（31%）**	82（9%）

（15）a. I don't want to ***see any woman be misdiagnosed***…　（COCA: 1990. SPOK）
b. I'd hate to ***see them be slimmed*** down into fewer.　（COCA: 2012. BLOG）
c. I'd like to ***see writers get paid*** fairly…　（COCA: 2012. BLOG）
d. I hate to ***see people get sucked*** into the idea…　（COCA: 1993. NEWS）

この調査から得られた結果は（16）のようにまとめられる。

（16）a. be -en 補文は，若干ではあるが，英米に差が見られた。　（cf. 表 4）
b. 先行研究で容認されていた get -en 補文は主にアメリカ英語に見られた。（cf. 表 4）
c. 先行研究の見解と異なり，現在完了形の知覚動詞が用いられている例はかなり限られていたが，感情を表す語句との共起が顕著であった。（cf. 表 6）

5．調査結果の考察

前節にて，BNC と COCA より得られたデータについてまとめ，英米における頻度の差，そして，使用される環境について分析を行った。本節では，更に考察を深めていく。まず, 英米に見られる受身表現の分布（=16a-b）について，Felser（1999: 83）もまたイギリス人英語母語話者は，be -en 補文の使用を避ける傾向にあるとする一方で，アメリカ人英語母語話者の多くは違和感を覚えることもあれば，文法的とみなす者もいるという。ではなぜ，英米において be -en 補文の使用頻度に差が見られるのだろうか。本論考では，get -en 補文の定着が，be -en 補文の容認可否性に影響を及ぼしている可能性について議論していく。まず，get 受動文について，Sussex（1982: 90）によればアメリカ英語に多く確認されるという。また，松元（2011: 22）によれば，20 世紀には主として米語の口語的表現において大量に用いられ，Schwarz（2017）および（2019）

は，20 世紀後半になると書き言葉においても劇的に増加するという。アメリカ英語の知覚動詞補文も同様に，get -en 補文が用いられ，その用法が確立すると，それに伴い，それまで非文法的とみなされていた be -en 補文もまた類似する表現として徐々に用いられるようになったと考えられる。このような事情から，アメリカ英語では，see NP {be / get} -en もまたイギリス英語よりも多く用いられていると考えられる。これに関連して，COHA を用いて調査を行ったところ，表 7 に示すように，あまり有意義な結果は得られなかったが，20 世紀後半から get + -en 補文が優勢となっている。

表 7. COHA における “{see / watch / hear} NP {be / get} + -en” の分布

	ALL	1830	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
be	54	1	1	2	1	4	1		3	5	1		2	3	2	4	8	11	5
get	160			1				1	3	2	8	7	**17**	6	**13**	**14**	**27**	**29**	**32**

さらに，COCA で検出されたデータを 10 年毎に分類した結果，表 8 に示すように，2010 年代から，be -en 補文と get -en 補文の総数が急激に増加していた。

表 8. COCA における “{see / watch / hear} NP {be / get} + -en” の分布

	1990s	2000s	**2010s**	TOTAL
be -en	62	52	**221**	335
get -en	187	196	**496**	879

またこれらの例をジャンル別に分析を行ったところ，表 9 に示すように，be -en 補文は SPOKEN に多く見られる一方で，get -en 補文は，表 10 のように ACADEMIC を除いて，ジャンルに偏ることはなく検出された。

表 9. COCA における “{see / watch / hear} NP be + -en” のジャンル別分布

	1990s	2000s	2010s	TOTAL
SPOKEN	24	13	38	75
NEWSPAPER	12	5	10	27
TV	12	8	12	32
MOVIE	7	7	11	25
FICTION	5	9	6	20
MAGAZINE	1	9	14	24
ACADEMIC	1	1	3	5
BLOG	0	0	67	67
WEB	0	0	60	60

表 10. COCA における “{see / watch / hear} NP get + -en” のジャンル別分布

	1990s	2000s	2010s	TOTAL
SPOKEN	25	26	37	88
NEWSPAPER	26	32	30	88
TV	49	53	64	166
MOVIE	43	42	65	150
FICTION	26	25	29	80
MAGAZINE	15	15	35	65
ACADEMIC	3	3	2	8
BLOG	0	0	139	139
WEB	0	0	95	95

これらの調査結果から，see NP be -en は元来口語で用いられていたものが，see NP get -en の使用拡大に伴い，徐々に書き言葉として用いられるようになり，増加した可能性が考えられる。また，（16c）の調査結果について，元来 be 受動態は当該の出来事が生じることを客観的に述べるのに対し，get 受動態は，話者やその他の関係者にとって不利益（ときには利益）となることを表すとされている（cf. Huddleston and Pullum（2002: 1442））[7]。COCA の知覚動詞補文においても，1990 年代から 2000 年代にかけて，感情表現と共起する be -en 補文は少数であったが，表 11 に示すように，get -en 補文の増加と共に 2010 年代になるとかなり増加していることがわかる。これらの調査結果から，be -en 補文

は get -en 補文の増加に伴い，類似する表現として徐々に用いられるようになったと考えられる。

表 11．COCA における {see / watch / hear} NP {be / get} +-en と共起する感情表現の分布

	1990s	2000s	**2010s**	TOTAL
感情表現＋ be -en 補文	10	7	**56**	73
感情表現＋ get -en 補文	58	72	**140**	270

6．まとめ

これまでの先行研究では，知覚動詞補文における be -en 補文は非文法的である一方で，それを進行形にした being -en 補文や動作受身を表す get -en 補文は文法的であるとされてきた。このことについて，BNC や COCA を用いて調査した結果，イギリス英語においては，be -en 補文は殆ど検出されず，アメリカ英語では僅かではあるが検出された。また先行研究で容認されている get -en 補文に関してもイギリス英語ではほとんど検出されず，アメリカ英語に多く検出された。このようにアメリカ英語において，それまで一般的に非文法的であるとされていた be -en 補文が容認されつつある要因として，get -en 補文の拡大が考えられる。元来 get 受動文はアメリカ英語の口語的表現であり，時代を経るにつれて書き言葉においても確立した表現として用いられているが，知覚動詞補文においても，get -en 補文が確立すると，それまで非文法的と見なされてきた be -en 補文に影響を及ぼし，徐々にアメリカ英語で be -en 補文が用いられるようになったと考えられる。

NOTES

（1）be -en 補文は To ***see her Coronation be performed***.（Shakespeare, *2 Henry Ⅳ*. 1.1.47）のように通時的に観察され，韻律などの影響を受けている可能性が考えられるが，EEBO を用いて調査したところ，表 i のように，being -en よりも be -en が優勢であった。

表ⅰ．EEBO における see NP {be / being} -en の分布

	15c	16c	17c	TOTAL
see NP be -en	3	32	161	196
see NP being -en	0	9	32	41

後述するように現代英語において be -en 補文が非文法的と見なされる要因は，原形不定詞が持つ完結性のアスペクトと be の状態性にあると考えられる。一方，知覚動詞補文に出現する準動詞の通時的調査を行った村岡（2021）によれば，初期近代英語において，準動詞のアスペクトの差は曖昧であったという。そのため，近代英語においては現代英語と異なり，see NP be -en のような表現は容認されていたと考えられる。

（2）村田･成田（1996: 133）によれば, be を用いた受身は用いられる動詞によって動作を表す動作受動と動作の結果としての状態を表す状態受動とで曖昧性が生じるという。さらに村田・成田（1996: 134）は be 受身では動作受動と状態受動とで曖昧になるのに対して，get 受身では動作受動でのみ用いられ，be 受身のような曖昧性は解消されるという。

（3）同様のことが，過去の習慣を表す would にも見られ，一般的に状態動詞と共起しない。これに関して，Declerck（1991）は（ⅱ）に示すように一定の期間における反復的な出来事を表す場合には，would は状態動詞とも共起できるという。これは個々の状態的な出来事が，繰り返しの動作として捉えられ，live が stay のような動作的な意味を帯びる為である。

（ⅰ）a. I {used to / ****would*****} *be* a waiter, but now I’m a taxi-driver.
（Alexander 1988: 235）
b. I {used to / ****would*****} *have* an old Rolls-Royce.　（Swan 2005[3]: 623）

（ⅱ）He ***would live*** at the Savoy whenever he came back to England.
（Declerck 1991: 417）

（4）しかし，Gee（1975）は，（ⅰ）の例を挙げ，（10）が「完了」の読みになり得る可能性を否定している。Gee（1975: 377）によれば，「ある特定の場面で

実際に（目の前で）知覚した出来事」を表す場合には，I ***saw a car be wrecked*** by the police. のような例は容認されないという。そして，（ia）を「完了･結果」用法と解釈できる場合には，過去形の saw とほぼ同じであり，「ある特定の場面で実際に（目の前で）知覚した出来事」を表すため容認されないが，（ib）のように補文主語を複数形にして，現在完了形を「経験」用法と解釈できる場合には「異なる場面で繰り返された出来事」を表すため，容認されるという。

（ｉ）a. I***'ve seen a car (#be) wrecked*** by the police. （Gee 1975: 377）
b. I***'ve seen cars be wrecked*** by the police. （ibid.）

（5）分析対象となる知覚動詞は see，watch と hear に限る。知覚動詞 taste と smell については，人間の五感で味覚と嗅覚は視覚・聴覚・触覚と比較すると感覚の精度が劣っており，外界の有界的出来事についての知覚をもたらし難いため，taste と smell は原形不定詞を補文に取れないとされている（cf. Pizer（1994: 340），Egan（2007: 146））。さらに，feel に関して，江川（1991[3]: 334）によれば，完結・非完結の区別が問題にならず，どちらも同じように用いられるとされることから，これらの動詞を本研究の対象から除外した。

（6）また (re)married，(un)dressed，rid 等の受身を表さない過去分詞は除外して調査を行った。

（7）影山（2007: 75-6）によれば，be には主観的なニュアンスは生じないが，get 受け身文は話者の主観的な受け止め方を表し，比率として迷惑の例がはるかに多いという。

参考文献

Akmajian, A. 1977. “The Complement Structure of Perception Verbs in an Autonomous Syntax Framework.” In Culicover, A, W., A. Akmajian and T. Wasaw. (Eds.), *Formal Syntax*. New York: Academic Press, 427-60.
Alexander, L. 1988. *Longman English Grammar*. London: Longman.
Allen, W. S. 1974[4]. *Living English Structure*. London: Longman.
安藤貞雄. 2005.『現代英文法講義』東京：開拓社.
安藤貞雄. 2008.『英語の文型 文型がわかれば，英語がわかる』東京：開拓社.
Arimoto, M. 1989. “Against the Raising Analysis of BE.” *English Linguistics* 6, 111-129.

Basilico, D. 2003. “The Topic of Small Clauses.” *Linguistic Inquiry* 34, 1-35.
Bolinger, D. 1974. “Concept and Percept: Two Infinitive Constructions and Their Vicissitude.” *World Papers in Phonetics Festschrift for Dr. Onishi's Kiju*. The Phonetic Society of Japan, 65-91.
Burzio, L. 1986. *Italian Syntax*. Dordrecht: Reidel.
Carlson, G. 1977. *Reference to Kinds in English*. Doctoral Dissertation. MIT.
Clark, R. and G. Jäger. 2000. “A Categorial Syntax for Verbs of Perception. *University of Pennsylvania Working Papers in Linguistics,*” Vol. 6: Iss. 3, Article 5, 15-33.
Declerck, R. 1981. “On the Role of Progressive Aspect in Nonfinite Perception Verb Complements.” *Glossa* 15, 83-114.
Declerck, R. 1991. *A Comprehensive Descriptive Grammar of English*. Tokyo: Kaitakusha.
Dixon, R. M. W. 2005[2]. *A Semantic Approach to English Grammar*. Oxford: Oxford University Press.
Egan, T. 2008. *Non-finite Complementation: A Usage-based Study of Infinitive and -ing Clauses in English*. New York: Rodopi.
江川泰一郎. 1991[3].『英文法解説 改訂三版』東京：金子書房.
Felser, C. 1999. *Verbal Complement Clauses: A Minimalist Study of Direct Perception Construction.* Amsterdam: John Benjamins Publishing Company.
Gee, J. P. 1975. *Perception, Intentionality, and Naked Infinitives: A Study in Linguistics and Philosophy*. Doctoral dissertation, Stanford University.
Gisborne, N. 2010. *The Event Structure of Perception Verbs*. Oxford: Oxford University Press.
Guasti, M. T. 1993. *Causative and Perception Verbs: A Comparative Study*. Torino: Rosenberg and Sellier.
Huddleston, R. and G. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
影山太郎. 2007.「get 受身文の統語構造と概念構造について」『英米文学』51(2), 63-82.
柏野健次. 1993.『意味論から見た語法』東京：研究社.
吉良文孝. 2006.「知覚動詞補文のアスペクト」『英語語法文法研究』13, 35-50.
Langacker, R. W. 2008. *Cognitive Grammar: A Basic Introduction*. Oxford University Press, Oxford.
Lapointe, S. G. 1980. “A Note on Akmajian, Steel and Wasow's Treatment of Certain Verb Complement Types.” *Linguistic Inquiry* 11, 770-87.
松元浩一. 2011.「18 世紀英語の get- 受動文」『長崎大学教育学部紀要：人文科学』77, 21-35.
Miller, G. 2002. *Nonfinite Structures in Theory and Change*. Oxford: Oxford University Press.
村岡宗一郎. 2021.「知覚動詞と使役動詞補文に出現する準動詞がもつアスペクト特性の発現時期について」近代英語協会第 38 回大会発表資料.
村田勇三郎・成田圭市. 1996.『テイクオフ英語学シリーズ 2 英語の文法』東京：大修館書店.

中右実. 1980.「テンス，アスペクトの比較」國廣哲彌（編）『日英語比較講座第 2 巻 文法』東京：大修館書店.

Palmer, F. R. 1965. *A Linguistic Study of English Verb*. London: Longman.

Palmer, F. R. 1974. *English Verb*. London: Longman.

Palmer, F. R. 1987[2]. *English Verb*. London: Longman.

Pizer, K. 1994. “Perception Verb Complementation: A Construction-based Account.” *Chicago Linguistic Society* 30, 335-346.

Schwarz, S. 2017. “Like Getting Nibbled to Death by a Duck: Grammaticalization of the Get-passive in the TIME Magazine Corpus.” *English Word-Wide* 37(3), 305-335.

Schwarz, S. 2019. “Signs of Grammaticalization. Tracking the Get-passive through COHA.” In Claridge, S. and B. Bös. (Eds.), *Developments in English Historical Morpho-Syntax*. Amsterdam: John Benjamins, 199-222.

Sheehan, M. and S. Cyrino. 2018. “Why Do Some ECM Verbs Resist Passivisation? A Phase-Based Explanation.” *Proceedings of the 48th Meeting of the North Eastern Linguistic Society* 48, 81-90.

白井賢一. 1999.「英語の知覚動詞構文の意味分析：認知意味論と形式意味論の「橋渡し」を目指して」『中京大学教養論叢』40(1), 1-61.

Sussex, R. 1982. “A Note on the Get-passive Construction.” *Australian Journal of Linguistics* 2, 83-95.

Swan, M. 2005[3]. *Practical English Usage*. London: Oxford University Press.

Wilder, C. 1992. “Small Clauses and Related Objects.” *Groninger Arbeiten zur germanistischen Linguistik* 34, 215-36.

（村岡宗一郎　日本大学大学院　Email: hollow_t_classic@ezweb.ne.jp）

英語コーパス学会　第47回大会

日　　時　　2021年10月２日（土）オンライン学会
開 会 式　　9:50–10:00

基調講演Ⅰ　10:00–11:00
認知意味論研究におけるコーパスと実験の利点と限界
松本　　曜（国立国語研究所）
司会・指定討論者　　森下　裕三（環太平洋大学）

〈Session 1　英語学・英文学（発表室１）〉
司　会　　杉森　直樹（立命館大学）

研究発表１　11:05–11:25
知覚動詞補文に出現する受身表現の容認可否について
村岡宗一郎（日本大学大学院）

研究発表２　11:30–11:50
中英語期の補部と２人称代名詞の構文関係：ICAMET による分析
泉類　尚輝（慶應義塾大学大学院）

研究発表３　11:55–12:15
多様な指標を組み込んだトピックモデル可視化ツールの開発とテクスト分析への応用
黒田　絢香（大阪大学大学院）

〈Session 2　英語教育（発表室２）〉
司　会　　堀家　利沙（神戸大学大学院）

研究発表１　11:05–11:25
日本人大学生 EFL 学習者の make + 名詞のコロケーション使用について
澤口　　遼（甲南高等学校・中学校）

研究発表２　11:30–11:50
Verification of the Effectiveness of 20 Months of Speaking Lessons for High School Learners: An Analysis of Fluency on the APTIS Speaking Test
Maxim TIKHONENKO（Tokyo University of Foreign Studies）
Keiko MOCHIZUKI（Tokyo University of Foreign Studies）

研究発表３　11:55–12:15

Learner Corpus-Based Study of L1 Effects on L2 English Auxiliary Verb Use: The Case of “Will”

Laurence NEWBERY-PAYTON（Tokyo University of Foreign Studies）

〈Session 3　言語資源開発（発表室３）〉

司　会　佐々木恭子（神戸大学大学院）

研究発表１　11:05–11:25

「1961-2021 日本語小説コーパス」の構築：
日英マインドスケープ対照研究の新しい可能性　石川慎一郎（神戸大学）

研究発表２　11:30–11:50

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

仁科　恭徳（神戸学院大学）
赤瀬川史朗（Lago 言語研究所）

研究発表３　11:55–12:15

授業コーパス構築のための自動タグ付けツール “Classroom Corpus Tagger” の開発

大橋由紀子（ヤマザキ動物看護大学）
片桐　徳昭（北海道教育大学）
押切　孝雄（戸板女子短期大学）

総会 12:15–12:40

〈Session 4　英語教育（発表室１）〉

司　会　和泉　絵美（京都大学）

研究発表１　13:30–13:50

日本人英語学習者の英語原因表現使用：ICNALE に基づく量的概観の新しい可能性

佐々木恭子（神戸大学大学院）

研究発表２　13:55–14:15

CEFR 準拠教科書における英語コロケーションの難易度変化要因の特定

畔元里沙子（九州大学大学院）

研究発表３　14:20–14:40

高校英語指導における句動詞の扱い－教科書とセンター試験の分析から－

堀家　利沙（神戸大学大学院）

〈Session 5　英語教育・英語学（発表室２）〉

司　会　　梶山　達也（同志社大学大学院）

研究発表１　　13:30–13:50

ライティング評価とテクストの特徴との相関関係：メタ分析による研究成果の統合

小島ますみ（岐阜女子短大）・金田　拓也（帝京大学）

研究発表２　　13:55–14:15

N-grams at the Beginning of the Moves in the Results Section of Experiment Medical Research Articles　　Tatsuya ISHII（Kobe City College of Technology）

Takeshi KAWAMOTO（Hiroshima University）

研究発表３　　14:20–14:40

オックスフォード・ユニオンにおけるリーダーシップ育成の示唆：英語圏のリーダーの発話コーパス分析　　中谷　安男（法政大学）

〈Session 6　ESP（発表室３）〉

司　会　　村岡宗一郎（日本大学大学院）

研究発表１　　13:30–13:50

工学系大学院生のための教材開発　－日英コーパスの分析

石川　有香（名古屋工業大学）

研究発表２　　13:55–14:15

生化学英語学術論文のための学術語彙リスト　　清水　　眞（東京理科大学）

村田　真樹（鳥取大学）

研究発表３　14:20–14:40

金融関連辞典と実務資料コーパスを用いた経済・金融分野の英語語彙リスト研究

小谷　尚子（東京外国語大学大学院）・佐野　　洋（東京外国語大学）

〈Session 7　英語学・英文学（発表室 1）〉

司　会　　田畑　智司（大阪大学）

研究発表１　　14:50–15:10

LDA Topic Modelling of Tennyson’s Poetry

Iku FUJITA（Osaka University Graduate School）

研究発表２　　15:15–15:35

チャンクのコロケーション：spaCy を用いた共起分析の試み

内田　　諭（九州大学）

研究発表３　15:40–16:00
動詞の意味はトピックから推測できるか：英語の動詞 run を例に
木山　直毅（北九州市立大学）・渋谷　良方（金沢大学）

〈Session 8　文法・統語（発表室２）〉
司　会　泉類　尚貴（慶應義塾大学大学院）

研究発表１　14:50–15:10
have long V-ed 構文の典型例　松田　佑治（立命館大学）

研究発表２　15:15–15:35
Distribution of Repeated Appearance of Grammar Items in Junior High School Textbooks through Nonlinear Regression　Yasuo AMMA（Dokkyo University）

研究発表３　15:40–16:00
現代スペイン語における主語後置の数理モデル化
小林純一朗（東京外国語大学）・佐野　洋（東京外国語大学）

〈Session 9　DDL（発表室 3）〉
司　会　村岡宗一郎（日本大学大学院）

研究発表１　14:50–15:10
英語の動詞－名詞コロケーション学習に対する DDL の効果
佐竹　由帆（青山学院大学）

研究発表２　15:15–15:35
中学生のための Web 版 DDL 支援ツールの開発と活用
西垣知佳子（千葉大学）
赤瀬川史朗（Lago 言語研究所）
川名　隆行（千葉大学教育学部附属中学校）
中井　康平（千葉大学教育学部附属中学校）
見目　慎也（千葉大学教育学部附属中学校）
山﨑　達也（千葉大学教育学部附属中学校）

研究発表３　15:40–16:00
Introducing AntConc 4.00: A Fast, Powerful, and Easy-to-Use Corpus Analysis Tool for Small and Large-Scale Corpus Analysis and Data-Driven Learning
Laurence ANTHONY（Waseda University）

基調講演 II 16:10–17:10

Statistics and Data Visualization in Corpus Linguistics with #LancsBox

Vaclav BREZINA（Lancaster University, UK）

司会・指定討論者 宇佐美裕子（東海大学）

閉 会 式 17:10–17:30

基調講演 I

「認知意味論研究におけるコーパスと実験の利点と限界」

松本　曜先生（国立国語研究所）

司会・指定討論者　森下　裕三（環太平洋大学）

認知言語学の量的研究においては，コーパスと実験の両方が用いられるが，それぞれどのような利点と限界があるのだろうか。保田（2011）は，「犬」などの具体名詞の意味を特定する際のコーパスと実験の有効性を検討し，動作に関する意味要素を抽出するのにはコーパスが向いているが，外見的特徴を抽出するのにはコーパスよりも描画実験が有効であるとしている。

また，松本（2021）は，移動事象の言語表現の研究におけるコーパスと発話実験の有効性を比較し，総合的な言語使用の実態を知るためにはコーパスに強みがあるが，そこに含まれる文がどのような事象を描いているのか確定できないという限界があるとしている。一方，発話実験は特定の移動事象の描写を引き出せる点で有効だが，その結果がどこまで言語使用の代表的なものかについて検討が必要であるとしている。

一般に，コーパスに基づく研究では，言語化されていない情報を扱いにくいという課題がある。コーパスと実験の利点と限界を認識した上で研究を進める必要がある。

主要参考文献

松本　曜　2021.「移動表現の研究におけるコーパスと実験」篠原和子･宇野良子(編)『実験認知言語学の深化』287-309．ひつじ書房

保田　祥　2011.『名詞の百科事典的意味の抽出方法とその有用性：内省・描画実験・コーパス調査』博士論文，神戸大学 .

基調講演 II

Statistics and Data Visualization in Corpus Linguistics with #LancsBox

Vaclav BREZINA (Lancaster University, UK)

司会・指定討論者　宇佐美裕子（東海大学）

In an ideal world, theory and practice would go together hand in hand. Strong theoretical and methodological grounding of corpus linguistic research leads to robust results, which can be meaningfully applied in practice. In reality, however, we can often see a disconnect between

high theoretical requirements of current research and what corpus linguists can actually do in their studies using existing tools. For example, we can see a limited range of statistical measures used in corpus linguistics, which until fairly recently typically relied on the log likelihood measure and a couple of collocation statistics (t-score and MI score) precisely because these were easily available in existing tools.

Corpus linguistics as a versatile methodology of language analysis (McEnery & Hardie 2011) thus requires access to appropriate software tools. These need to be able to cope with increasing demands on the sophistication of the analysis and increasing size of the data. In recent years, researchers have been critically re-evaluating the existing procedures in the field and have proposed more rigorous approaches to data analysis (e.g. Kilgarriff 2005, 2012; Gries 2006, 2013; Lijffijt et al. 2014; Brezina & Meyerhoff 2014; Brezina et al. 2015; Gablasova et al. 2017; Brezina 2018). Reflecting on this debate and combining statistical sophistication and accessibility is the main challenge that needs to be met by corpus linguists today; the analyses should encourage a multi-dimensional view of data, easy comparison, and effective visualization.

In this lecture, I will deal with key questions of corpus methodology and statistics and the implementation of statistical solutions in the #LancsBox software (Brezina et al. 2015). #LancsBox is a free multi-platform tool, which can analyse any language. It can be used by linguists, language teachers, translators, historians, sociologists, educators and anyone interested in quantitative language analysis. Extensive documentation about #LancsBox is available, also in Japanese: http://corpora.lancs.ac.uk/lancsbox/docs/pdf/LancsBox_5.1_manualJP.pdf

References

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Brezina, V., & Meyerhoff, M. (2014). Significant or random. A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, *19*(1), 1–28.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139–173.

Gries, S. T. (2013). *Statistics for linguistics with R: a practical introduction*. Berlin: Walter de Gruyter.

Gries, S. Th. (2006). Some proposals towards a more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik*, *54*(2), 191–202.

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus linguistics and linguistic theory*, *1*(2), 263–276.

Kilgarriff, A. (2012). Getting to know your corpus. In Sojka, P., Horák, A., Kopecek, I. & Pala,

K. *Text, Speech and Dialogue* (pp. 3–15). Berlin: Springer.
Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K., & Mannila, H. (2014). Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, advanced access.
McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

【研究発表 Session 1　英語学・英文学】
【研究発表１】

知覚動詞補文に出現する受身表現の容認可否について

村岡宗一郎（日本大学大学院生）

現代英語の知覚動詞は補文に原形不定詞や現在･過去分詞をとるが, Bolinger（1974）などの先行研究によれば，I saw the children be beaten. のような受身表現は一般的に容認されない。しかし，村上春樹の『騎士団長殺し』の英訳には，I couldn't stand to see her be cremated. という例が確認される。Palmer（1968）はこのような例は稀であると分析をするが，Palmer（1974）や（1987）からその記述は削除されている。また，多くの先行研究では，I saw the children get beaten. は容認されるというが，BNC と COCA を用いて調査を行った結果，このような例はアメリカ英語にのみ用いられていることが明らかになった。本研究では，BNC や COCA を用いて，英米における使用頻度から知覚動詞補文に出現する受身表現の容認可否について分析していく。

主要参考文献

Bolinger, D. 1974. “Concept and percept: Two infinitive constructions and their vicissitude.” *World Papers in Phonetics Festschrift for Dr. Onishi's Kiju*, 65-91. The Phonetic Society of Japan.
Palmer, F, R. 1968. *Linguistic Study on the English Verb*. London: Longman.
Palmer, F, R. 1987. *The English Verb*, 2nd Edition. London: Longman.

【研究発表２】

中英語期の補部と 2 人称代名詞の構文関係：ICAMET による分析

泉類　尚貴（慶應義塾大学大学院生）

英語史における変化の一つに, 補部の変化がある。定型節から非定型節への変化は,

中英語期にその萌芽が見られる（Los, 2005）。補部と意味の関係について，Rohdenburg（1995）によれば，非定形節のほうが coercive force が強いことが示されている。Coercive force の強さは，command をはじめとする指令動詞の分析から提示された。一方で，同一の動詞が異なる補部を従える例も見られる。本研究では，補文の変化が起こりだした時代である中英語期に焦点をあてて，指令動詞の現れる構文の一つである，明示的遂行文における 2 人称代名詞の用法（thou 系か ye 系か）と補部の that 節，不定詞の構文関係について，Innsbruck Computer Archive of Machine-Readable English Texts（ICAMET）を中心に収集したデータを示し，分析の可能性を示す。

主要参考文献

Los, Bettelou. *The Rise of the To-Infinitive*. Oxford: OUP, 2005.

Manabe, Kazumi. *The Syntactic and Stylistic Development of the Infinitive in Middle English*. Kyushu University Press, 1989.

Rohdenburg, G. 'On the Replacement of Finite Complement Clauses by Infinitives in English.' *English Studies*. 76:4 (1995), 367-388.

【研究発表３】

多様な指標を組み込んだトピックモデル可視化ツールの開発とテクスト分析への応用

黒田　絢香（大阪大学大学院生）

機械学習アルゴリズムの一つであるトピックモデルを用いたテクスト分析において，各トピックの特徴や関係性，構成単語や出現傾向など様々な要素を的確に可視化し，モデルの全体像を把握することが極めて重要である．本研究では，従来の分析で主に用いられていた「文書ごとのトピック出現確率」「各単語の weight」に加えて，coherence や exclusivity，effective number of words など様々な指標を組み込んだビジュアライゼーションツールを開発し，いかに効果的にトピックを可視化できるか，それらがどのように文学作品分析に寄与するかを検討する．

主要参考文献

Jockers, M. and Mimno, D.: Significant themes in 19th-century literature. *Poetics* 41: 750–769 (2013)

田畑智司 : FLOB コーパスの意味構造 : 確率論的トピックモデルによる言語使用域の特徴付け『統計数理研究所 共同研究リポート』386: 1–17 (2017)

【研究発表 Session 2　英語教育】
【研究発表１】

日本人大学生 EFL 学習者の make ＋名詞のコロケーション使用について

澤口　　遼（甲南高等学校・中学校（非））

本研究の目的は，学習者の同一トピックについての英日のエッセイを集めた学習者コーパスである KUBEC を用い，日本人大学生 EFL 学習者の英語動詞 make ＋名詞のコロケーション使用において，日本語「つくる」の意味体系が与える影響が習熟度と共にどのように変化するか調査することにある。

母語の影響を受けていると考えられるコロケーション使用率の頻度を学習者間で比較した結果，習熟度別では差が見られず，全習熟度の学習者が「関係」や「環境」など, make よりも抽象的な名詞と共起する「つくる」の多義性の影響を受けたコロケーションを使用していることが明らかになった。

これらの結果から，学習者へのコロケーション指導への示唆を考察する。

主要参考文献

山西博之・水本篤・染谷泰正.（2013）.「関西大学バイリンガルエッセイコーパスプロジェクト―その概要と教育研究への応用に関する展望―」.『関西大学外国語学部紀要』9, 117-139.

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics, 22*(2), 173–195. doi: https://doi.org/10.1093 /applin/22.2.173.

Anthony, L. (2017). AntPConc (Version 1.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.html

【研究発表２】

Verification of the Effectiveness of 20 Months of Speaking Lessons for High School Learners: An Analysis of Fluency on the Aptis Speaking Test

Maxim TIKHONENKO（Tokyo University of Foreign Studies）
Keiko MOCHIZUKI（Tokyo University of Foreign Studies）

We present a 20-month longitudinal study on the development of speaking ability among Japanese high school learners of English who participated in online speaking lessons. Students were divided into an experimental group of 32 students who took 20 monthly lessons and a

control group of 22 students who took 3 lessons.

Both groups then took the Aptis speaking test three months after the last speaking lesson, the fall semester of the third year. Speaking data recorded from the test was transcribed using ELAN, and the durations of pauses and speaking time were measured. The transcribed data was then divided into AS-Units and analyzed from the perspectives of fluency and complexity.

For fluency, speech rate, ratio of pause time to speaking time, number of pauses per minute, ratios of self-corrections, repetitions, and fillers to AS-Unit were measured. The analysis showed that speech rate and the ratio of pause time to speaking time ratio showed the greatest difference. For complexity, the ratio of subordinate clauses to AS-Units and the mean number of words were measured. These measures differed less than the fluency measures between the two groups.

The analysis showed that the experimental group was significantly more fluent than the control group.

References

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: OUP.

Foster, P., Tonkyn, A., and Wigglesworth, G. (2000). "Measuring spoken language: A unit for all reasons", *Applied Linguistics*, *21*(3), 354–375, Oxford: Oxford University Press.

Housen, A., Kuiken, F., Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In Housen, A., Kuiken, F., Vedder, I. (Ed.). *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Language Learning & Language Teaching 32, pp. 1–20.

【研究発表３】

A Learner Corpus-Based Study of L1 Effects on L2 English Auxiliary Verb Use: The Case of "Will"

Laurence NEWBERY-PAYTON（Tokyo University of Foreign Studies）

This study analyzes use of the modal verb "will" by L1 Chinese and Japanese learners of English. Data is drawn from the Written Essays module of the International Corpus Network of Asian Learners of English (ICNALE). Both frequency and type of use are predicted to be influenced by the presence (in Chinese) or absence (in Japanese) of functional equivalents to "will" in L1. Analysis reveals overuse of "will" by L1 Chinese learners across proficiency levels. This high frequency of use can partially be attributed to Chinese learners' consistent use of "will" to express non-future (e.g. habitual or generic) meanings. Such uses are analogous to functions of the modal auxiliary "hui" in Chinese, suggesting potential L1 influence. In

contrast, Japanese learners not only use “will” less frequently, but also consistently omit it in obligatory contexts. The two groups of learners also differ in their use of “will” in conditional sentences. With rising proficiency, “will” becomes increasingly restricted to conditional sentences in essays by Japanese learners, whereas the opposite trend is observed among Chinese learners. Finally, the study considers task-related effects, notably that the convergence on native speaker-like frequency of use is apparent in only one of the two essay topics.

References

Bardovi-Harlig, K. (2017). Beyond individual form-meaning associations in L2 tense-mood-aspect research. In M. Howard & P. Leclercq (Eds.), *Tense-aspect-modality in a second language contemporary perspectives* (pp. 27–52). Amsterdam: John Benjamins.

Nakayama, S. (2021). Contrastive interlanguage analysis of Japanese EFL learners' modal auxiliary verb use in conversation. *Journal of Educational Research and Review*, *4*(1), 1–13.

Tsai, W. (2015). On the topography of Chinese modals. In U. Shlonsky, (Ed.), *Beyond functional sequence* (pp. 275–294). Oxford: Oxford University Press.

【研究発表 Session 3　英語教育】
【研究発表 1 】

「1961-2021 日本語小説コーパス」の構築：
日英マインドスケープ対照研究の新しい可能性

石川慎一郎（神戸大学）

構築中の日本語小説コーパスについて報告する。これは Brown Corpus の標本抽出年である 1961 年を起点として，2021 年まで，10 年ごとの間隔で 3 大文芸誌（「新潮」「文學界」「群像」）に掲載された日本語小説とその英訳（機械翻訳 2 種）を収集するものである。本コーパスを用いることで，ジャンル要因を統制した上で，現代日本語の経年変化を調査することができる。また，付随する英訳データを Brown/LOB（1961 年），Frown/FLOB（1991-92 年），Crown/CLOB（2009 年）等の小説データと対照することで，時代要因を統制した上で，日英小説の言語・文体・マインドスケープの比較を行うこともできる。発表では本コーパスの開発理念と手順，また，収集済みのデータから明らかになった知見の一部を報告する。

主要参考文献

石川慎一郎．（2015）．「FROWN/FLOB Corpus および BCCWJ データの再構成に基づく英日対照言語研究用小説テキストデータセットの構築の試み：English-

Japanese Modern Fiction Corpus (EJ-MoFic) の概要」『統計数理研究所共同研究レポート』*340*, 1–18.
石川慎一郎.（2021）.『ベーシックコーパス言語学』第 2 版 . ひつじ書房.
三竹保宏.（2018）.「Deep Learning による AI 機械翻訳のイノベーション」『ビジネスコミュニケーション』*55*(8), 11.

【研究発表２】

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

仁科　恭徳（神戸学院大学）
赤瀬川史朗（Lago 言語研究所）

本発表では，我々が現在開発中の網羅型日英・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）について中間報告を行う。まず，現在までに構築された日英・英日パラレルコーパスや検索ツール，それらを活用した研究を振り返り，今後の展望を述べる。次に，一般参照パラレルコーパスの構築に先駆けて，実験的に開発している『（仮称）パラレルリンク』（Ver. 1.0）の開発状況等を報告する。特に，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オンライン上で網羅的に串刺し検索できる本検索システム（Ver. 1.0）は，擬似的な一般参照パラレルコーパスとして活用することができ，検索語に関する精緻な語彙プロファイルの獲得やオーセンティックな翻訳例の獲得において有益なツールとなる可能性があることも示唆したい。

主要参考文献

中條清美・西垣知佳子・赤瀬川史郎・内山将夫.（2015）.「レキシカル・プロファイリング型オンラインコーパス検索ツール LWP for ParaNews の英語授業における利用」『日本大学生産工学部研究報告 B』第 48 号，45–57.
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.
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.

【研究発表３】

授業コーパス構築のための自動タグ付けツール “Classroom Corpus Tagger” の開発

大橋由紀子（ヤマザキ動物看護大学）
片桐　徳昭（北海道教育大学）
押切　孝雄（戸板女子短期大学）

本研究では，文字起こしされた各発話に対してタグを自動生成する Classroom Corpus Tagger（CCT）を紹介する。授業コーパス構築には，話者・使用言語・活動等に関するタグ付与を要する（e.g., Ohashi & Katagiri, 2016）。手作業でのタグ付与は時間を要し，ミスが生じやすいことが課題であった。CCT は JavaScript を利用，ブラウザで作動し，日本語か英語を自動的に判別して言語タグを付与する。話者タグは任意に複数種類の設定が可能である。これにより授業コーパス構築が容易となる。手動と CCT でのタグ付けの妥当性の比較実験を行った結果，手動構築で見られるミスは，CCT を利用した場合は見られず，CCT を使用したタグ付けの正確性と，負担軽減が観察できた。発表では，CCT の実演と実験結果の詳細について報告する。

主要参考文献

Ohashi, Y., & Katagiri, N. (2016). The effects of explicit instructions observed in teacher transcripts and student impression remarks in elementary school. *HELES Journal* (16), 3–18.

【研究発表 Session 4　英語教育】
【研究発表１】

日本人学習者の英語原因表現使用：ICNALE に基づく量的概観

佐々木恭子（神戸大学大学院生）

日本人学習者は英作文において because を多用するとされるが（小林，2009；佐々木，2021），幅広い原因表現を対象にした計量的検証は十分になされていない。そこで本研究は，Altenberg（1984），Biber et al. (1999) などから取捨選択した 34 種の原因表現を対象に，ICNALE の母語話者作文と，CEFR A2 学習者，B1_2/B2+ 学習者作文を比較した。その結果，（1）母語話者は接続詞と動詞で理由を深化する使用傾向，（2）学習者は接続詞と reason など名詞使用により理由を羅列的に表す傾向などの知見が得られた。本研究の知見は，高校段階での英語教育の改善に一定の示唆を有する。

主要参考文献

Altenberg, B. (1984). Causal linking in spoken and written English. *Studia Linguistica*, *38*(1) 20–69. http://dx.doi.org/10.1111/j.1467-9582.1984.tb00734.x

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson Education Ltd.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman Group Limited.

【研究発表２】

CEFR 準拠教科書における英語コロケーションの難易度変化要因の特定

畔元里沙子（九州大学大学院生）

学習者にとって初級の単語から成る英語コロケーションが必ずしも難易度が低いとは限らないように，英語コロケーションの難易度の測定は難しい（内田，2015）。そこで本研究は，英語コロケーションの難易度が変化する要因を客観的指標から特定することを目的とする。そのためにまず，自作の CEFR 準拠の教科書コーパスに含まれるレベル別サブコーパスを利用し，そこに一定の基準以上で出現する二語で構成されるコロケーションを抽出，コロケーションのレベル別リストを作成した。その後，統計的手法を用いてレベル間のコロケーションの特徴の違いを分析し，CEFR 準拠の教科書コーパス内でのコロケーションのレベル変化要因を明らかにした。

主要参考文献

内田諭．（2015）．基本動詞のコロケーション難易度測定―CEFR レベルに基づくテキストコーパスからのアプローチ―．言語処理学会年次大会発表論文集，*21*，880–883.

【研究発表３】

高校英語指導における句動詞の扱い―教科書とセンター試験の分析から―

堀家　利沙（神戸大学大学院生）

日本人高校生の英語アウトプットにおける句動詞運用を分析した研究は多いが，高校生向けインプット資料における句動詞の実態を計量的に調査した研究は必ずしも多くない。そこで，本研究は高校教科書 6 種，センター試験 10 年分をコーパス化し，BNC/COCA と比較した。句動詞の全体使用量，重要句動詞に対するカバー率，特徴

句動詞の 3 観点で調査した結果，（1）初中級教科書における頻度は母語話者の基準以下，（2）重要句動詞カバー率は，上級教科書，センター試験併用型であっても 7 割弱，（3）インプット資料の上級化に伴い，特徴句動詞の質も段階的に変化する，という 3 点が明らかになった。本研究の知見は，句動詞指導の改善に一定の意義を持つ。

主要参考文献

Gardner, D. & Davies, M. (2007). Pointing Out Frequent Phrasal Verbs: A Corpus-Based Analysis, *TESOL Quarterly, 41*(2), 339–359.

石井康毅．（2018）．「話し言葉コーパスと検定教科書に基づく日本人英語学習者の句動詞使用実態の分析」*Learner Corpus Studies in Asia and the World, 3*, 101–119.

石川慎一郎．（2019）．「英語教育における連語：ターゲット・インプット・アウトプットの三元コーパス分析をふまえた English N-gram List for Japanese Learners of English（ENL-J）の開発と利用」『言語分析のフロンティア』金星堂，32–47.

【研究発表 Session 5　英語教育・英語学】
【研究発表 1 】

ライティング評価とテクストの特徴との相関関係：メタ分析による研究成果の統合

小島ますみ（岐阜市立女子短期大学）
金田　　拓（帝京科学大学）

本研究は，小島・金田（2020）の研究対象を拡大し，2016 年以降の研究や 3 種類の調整変数を加えたものである。第二言語（L2）学習者のライティング評価とテクストの特徴の相関関係について，103 の研究（総参加者 15,537 人）のメタ分析を行った。結果より，ライティング評価と最も相関が高かったテクストの計量的特徴は流暢性であり，続いて正確性，語彙的複雑性，統語的複雑性，結束性の順であった。主観的な評価項目では，内容と言語使用の効果量が最も大きく，結束性，一貫性が最も小さい結果となった。また，書き手の年齢，母語，学習環境，ライティング評価方法，計量言語指標の種類が有意な調整変数であった。

主要参考文献

小島ますみ・金田拓．(2020)．「ライティング評価と言語的指標の関係―メタ分析による研究成果の統合」石井雄隆・近藤悠介（編著）．『英語教育における自動採点―現状と課題』(pp. 33–72）ひつじ書房

【研究発表 2】

N-grams at the Beginning of the Moves in the Results Section of Experimental Medical Research Articles

Tatsuya ISHII（Kobe City College of Technology）
Takeshi KAWAMOTO（Hiroshima University）

Using a corpus based on move analysis of experimental medical research articles (in total approximately 1.5 million words), Ishii & Kawamoto (2020) focused on the behavior of adverbs and successfully identified 26 lexical phrases for the three moves in the Results section: (RM1) Introducing experiments, (RM2) Announcing results, and (RM3) Commenting results. However, although there are cycles of the three moves, it was still unclear as to how the moves start and are connected. In this study, to identify the n-grams at the beginning of the three moves, we extracted and examined the first sentences of the three moves. After the first sentences of the three moves were extracted by CasualConc (2021) with the use of a wildcard, they were copied and pasted into Excel to divide them into independent words. The frequencies of the n-grams were counted with the help of CasualConc (2021). In conclusion, the observation of the n-grams led to the description of highly frequent phrases for starting and connecting moves; for example, the phrase to determine in (RM1), the phrase we observed in (RM2), and the phrase taken together these results in (RM3). This study will provide new insights for investigating a corpus based on move analysis.

References

Ishii, T., & Kawamo, T. (2020). The behavior of adverbs in the results sections of experimental medical research articles: A corpus-based move analysis. *English Corpus Studies, 27*, 23–52.

【研究発表３】

オックスフォード・ユニオンにおけるリーダーシップ育成の示唆：英語圏のリーダーの発話コーパス分析

中谷　安男（法政大学）

世界有数のディベート組織であるオクスフォード・ユニオンや，TEDTalk で講演を行った政治・経済界の代表者の発話コーパスを分析することにより，リーダーに必要な聴衆を説得する Communication Strategy（CS）の検証を行った。120 名の発話デー

タをテキスト化し約 50 万語のコーパスデータとして活用した。これを AntConc Windows (3.5.8) を使い FLOB 及び FROWN コーパスの合計 200 万語と比較し Keyword 分析により特徴語を抽出した。さらにこの特徴語のクラスター表現を抜きだし，リーダーたちが活用する CS を確認した。結果として，リーダーは聴衆の注意を喚起し，誘導し，積極的に行動を起こすように促す CS を効果的に活用していた。

主要参考文献

Charteris-Black, J. (2006). *The Communication of Leadership: The Design of Leadership Style*. London: Routledge.

Fetzer, A., and Bull, P. (2012). Doing leadership in political speech: Semantic processes and pragmatic inferences. *Discourse & Society, 23*(2), 127–144.

Kotter, J. P. (1999). *John P. Kotter on What Leaders Really Do*. MA: Harvard Business School Press.

【研究発表 Session 6　ESP】
【研究発表 1 】

工学系大学院生のための教材開発 －日英コーパスの分析

石川　有香（名古屋工業大学）

工学系大学では，大学院生に対しても，英語による論文執筆を求める声が強くなってきている。一方で，学部・大学院において，英語教育に割り当てられた時間は非常に少なく，英語論文執筆の体系的な指導はほとんどなされていない。工学系大学院生を対象とした教材開発が急務となっている。本プロジェクトでは，日英パラレルコーパスを作成し，日本語で研究活動を行っている工学系大学院生が，その研究成果を英語で発表するための教材開発を目指す。すでに，Academic Phrase Bank や AWSuM などすぐれたライティング支援ツールがあるが，英語使用に慣れていない工学系大学院生にはハードルが高くなっている。日本語の分析や日英の提示方法などの課題について考える。

主要参考文献

石川有香.（2021）.『ジャンルとしての工学英語』大学教育出版 .

【研究発表 2】

生化学英語学術論文のための学術語彙リスト

清水　　眞（東京理科大学）
村田　真樹（鳥取大学）

Shimizu et al. (2018) は，Hyland & Tse (2007) にならい，有機化学論文誌，物理化学論文誌に掲載された論文からコーパスを編纂し，有機化学論文のための学術語彙リスト，物理化学論文のための学術語彙リストを作成した。この研究では，2016 年と 2020 年に発行された生化学論文誌からコーパスを編纂し，生化学論文のための学術語彙リストをふたつ作成した。ふたつのリストの比較，ふたつのリストと有機化学，物理化学のリストとの比較を行った。どのリストにおいても，基礎単語は約 800 語しか用いられていないこと，基礎単語であっても，専門的な意味を持つものがあることなどがわかった。

主要参考文献

Hyland, K. & Tse, P. (2007). 'Is there an "Academic Vocabulary"?' *TESOL Quarterly* 41:2, 235–253.
石川 慎一郎. (2012). 『ベーシックコーパス言語学』 ひつじ書房.
Shimizu, M., Murata, M, & Ramonda, K. (2018), 'Teaching English for Chemistry at a Japanese University', *The Online Journal of Science and Technology* - July 2018 Volume 8, Issue 3

【研究発表３】

金融関連辞典と実務資料コーパスを用いた経済・金融分野の英語語彙リスト研究

小谷　尚子（東京外国語大学大学院生）
佐野　　洋（東京外国語大学）

本研究は，経済・金融分野を対象とした ESP 教育に関する教育的な観点からの語彙カテゴリ化と学習語彙のリスト化を検討している。発表では，辞書の見出し語における重要性と，実務文書における語の出現の有り様の二面から調査した結果を報告する。具体的には，英語及び日本語で出版されている経済・金融分野の用語辞書を対象に，それら辞書の見出し語を比較し，共通語彙を明らかにする。辞書の選択条件は，近年の出版であること，見出し語数が多いこと（5000 語以上），知名度が高い（流通量が多い）ことである。共通語彙の通用性を確認あるいは検証することを目的として，

主要米国企業 30 社の 10K（米国の有価証券報告書に該当）からテキストを取り出し（数字からなる語を含む 240 万語程度），使用されている語彙群をもとめ（170 万語程度），上記の共通語彙との重複や頻度傾向を調査した。経済・金融分野における実務上の観点からみた学習語彙のリスト化について，語彙の定性分析も含めて考察する。

主要参考文献

Dictionary of Finance and Banking（Oxford University Press, 2014）
『金融関連専門辞書経済・金融ビジネス英和大辞典』（日外アソシエーツ，2012）
Apple Inc. (2019). *Form 10-K 2018*. U.S. Securities and Exchange Commission.

【研究発表 Session 7　英語学・英文学】
【研究発表 1】

LDA Topic Modelling of Tennyson's Poetry

Iku FUJITA（Osaka University Graduate School）

Topic modelling is considered a promising approach in text mining (Meeks and Weingart, 2012), and a number of studies have examined prose texts using topic modelling (Tabata, 2017; Kiyama, 2018; Huang, 2020). However, the application of topic modelling to poetry is still developing; this study thus will make a step further towards an in-depth investigation using LDA (Blei et al., 2003) on Alfred Tennyson's poetry works.

The data of this study is a Victorian poet Alfred Tennyson's 66 epic and lyrical poems over 1,000 words in length. In this study, some explicit features to characterize works, such as character names and honorific titles, are excluded from the analysis as stopwords.

Emerging results LDA have shown the latent topics hidden behind prominent elements of poems in the corpus, and the topics appeared in some works in common; of further interest is the latent connections between some works. In addition, this study discusses the possibility of detecting rhyming elements when LDA is run on poetry data as well as the issue of PoS tagging on verse texts, suggested by the results of LDA in hindsight, and conceivable future approaches for addressing the issue.

References

Blei, M. D., Ng, Y. A., and Jordan, I. M. (2003). "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3, 2003, pp. 993–1022.
Meeks, E. and Weingart, B. S. (2012). "The Digital Humanities Contribution to Topic Modeling." *Journal of Digital Humanities*, Vol. 2, No. 1 Winter 2012, pp. 1–6.
Tabata, T. (2017). "Mapping Dickens's Novels in a Network of Words, Topics, and Texts: Topic

Modelling a Corpus of Classic Fiction." *Japanese Association for Digital Humanities Conference 2017, September 2017, Doshisha University*.

【研究発表２】

チャンクのコロケーション：spaCy を用いた共起分析の試み

内田　　諭（九州大学）

コロケーションの集計は多くの場合，単語単位で行われる。この集計方法の場合，higher education, the United States などのまとまりの情報（チャンク）が失われ，重要な共起情報を見逃してしまう可能性がある。本研究では，自然言語処理のアプリケーションを用いてチャンクの情報を保持した形でのコロケーションの集計を試みる。具体的には，Python のライブラリの一つである spaCy を用いて，名詞句のチャンクをタグ付けし，そのテキストデータに対して共起分析を行う。実施手順等を紹介した後，分析手法の妥当性，言語研究における有用性等について議論する。

【研究発表３】

動詞の意味はトピックから推測できるのか : 英語の動詞 run を例に

木山　直毅（北九州市立大学）
渋谷　良方（金沢大学)

近年，コーパス言語学では多義語の意味を調査する上でコロケーションや語が現れる統語や形態素といった文法要素や共起語の意味カテゴリーを用いる手法が提案されている（e.g. Gries 2006，Heylen et al. 2012）。本発表では，上述の要因に加え，語が現れるトピックが多義語の意味を決める要因になることを提案する。英語の動詞 run は非常に多くの意味を持つが（Gries 2006, Langacker 1988, Taylor 1996），例えばスポーツの話題で run が用いられれば「走る」の意味で，ビジネスの話題で run が現れれば「経営する」の意味だと考えるのが直感的には自然である。本発表では，News on the Web corpus (Davies 2016–) より抽出した run の事例に対し，トピックモデルの手法の 1 つである biterm topic model (Yan et al. 2013) を用いて以上の直感を実証する。

主要参考文献

Gries, S. (2006). Corpus-based Methods and Cognitive Semantics: The Many Senses of to run. In A. S. Gries Stefan (Ed.), *Corpora in Cognitive Linguistics Corpus-Based Approaches to Syntax and Lexis* (pp. 57–99). Mouton de Gruyter.

Davies, M. (2016). Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day.
Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A Biterm Topic Model for Short Texts. *Proceedings of the 22nd International Conference on World Wide Web*, 1445–1456.

【研究発表 Session 8　文法・統語】
【研究発表１】

have long V-ed 構文の典型例

松田　佑治（立命館大学）

住吉（2020）は，have long V-ed 構文の典型例を探るために，Web 版の COCA を対象とし，[have] long [VVN] という検索で，この構文の典型例を探り，複数の特徴を示している。しかし，住吉（2020）の調査で用いた品詞タグ [VVN] は，確かに過去分詞を抽出するものの，been を抽出しないという致命的問題などが複数見られる。そこで本発表では，COCA-FullText を対象とし，正規表現を用いて，品詞タグに頼らずに過去分詞を抽出するという新たな分析を行った。その結果，新しい知見の中の一つとして，have long V-ed 構文の V-ed には，圧倒的に been が生起することが分かった。

主要参考文献

Hilpert, Martin. 2014. *Construction Grammar and its Application to English*, Edinburgh: Edinburgh University Press.
住吉誠．2020.「コーパスと英語語法研究」『コーパス研究の展望（Aspects of English Corpus Studies)』（最新英語学・言語学シリーズ　第 11 巻）石川慎一郎・長谷部陽一郎・住吉誠（著）．東京 : 開拓社.
滝沢直宏．2016.「コーパスからの情報抽出と抽出データの意味づけに関わる諸問題」『英語コーパス研究』23: 45–60.

【研究発表２】

Distribution of Repeated Appearance of Grammar Items in Junior High School Textbooks through Nonlinear Regression

Kazuo AMMA（Dokkyo University）

Opportunities for repeated learning is of vital significance especially in second language acquisition. However, in the Japanese junior high school (JHS) context the beginner-level

grammar items are linearly arranged in the curriculum and occasional reviewing of past items seems to be neglected. This study is aimed at characterising the reappearance patterns of grammar items and differentiating them qualitatively, thereby suggesting the teacher's approach to individual items.

38 popular grammar items were selected for analysis in six MEXT-inspected JHS textbooks across three year grades (all published in 2016). The frequency was counted for each appearance printed in the student textbook as well as exercise answers and audio scripts in the teacher's manual; ie., for all exposures either visually or orally presented including repeated exercises.

A cumulative frequency data was collected to which a cubic regression was applied, resulting in high rates of squared residuals for high-frequency items (R2=0.95~0.99). The coefficients of the regression formulae were then used for factor analysis. The distribution of items revealed a new dimension of convex curve patterns and concave curve patterns, indicating how soon or late the items appear and reappear. It also showed textbook-specific patterns as well as universal ones.

References

Amma, K. (2018). Extracting patterns from transition of occurrence frequency of grammar items in a junior high school textbook. *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference (APCLC 2018)*, 219-226.

林正頼・石井康毅・高村大也・奥村学・投野由紀夫.（2016）. CEFR-based Coursebook Corpus からの CEFR レベル別基準特性の特定 . 投野由紀夫（代表）『学習者コーパスによる英語 CEFR レベル基準特性の特定と活用に関する総合的研究』（平成 24 年度～平成 27 年度科学研究費補助金（基板研究（A））研究課題番号 24242017 研究成果報告書）.

石井康毅.（2016）. CEFR-J Grammar Profile の構築のための英文法項目の選定・抽出・頻度集計・精度評価 . 投野由紀夫（代表）『学習者コーパスによる英語 CEFR レベル基準特性の特定と活用に関する総合的研究』（平成 24 年度～平成 27 年度科学研究費補助金（基板研究（A））研究課題番号 24242017 研究成果報告書）.

【研究発表３】

現代スペイン語における主語後置の数理モデル化

小林純一朗（東京外国語大学学部生）
佐野　　洋（東京外国語大学）

スペイン語の主語後置の発生メカニズムを多変量カテゴリの分類問題に帰着させ,

大規模データ（約 20 億語規模のコーパス）を用いて分類器（ロジスティック回帰モデル）を構築したので，本発表にて報告する。先行研究では，情報の新旧などの情報構造に基づいた手法が多く，情報の新旧などをカテゴリ変数とする線形重回帰分析もある。本研究では大規模データを高速分析すべく，表層に出現するカテゴリ変数（冠詞の定性など）を用いてモデルを作成した。その結果，先行研究で示された結果と同等かそれ以上の分類性能を持つことが分かった。スペイン語のみならず英語を含め，通言語的な情報構造概念が表層特徴から近似されうることが示された。

主要参考文献

Brunetti, L., &. Bott, S. (2011). Subject inversion in Romance: a corpus-based study; Handout distributed at: Quantitative Investigations in Theoretical Linguistics QITL-4.〈available in https://edoc.hu-berlin.de/bitstream/handle/18452/2021/brunetti.pdf?sequence=1〉

Hatcher, A. G. (1956). Theme and Underlying Question: Two Studies of Spanish Word Order. *Word* 12, supplement 3, 1-52.

Müller, A. C. & Guido, S. (2017). *Introduction to Machine Learning with Python* O'Reilly Media, Inc.,（アンドレアス・C・ミュラー & サラ・グイド　中田秀基（訳）.（2017）.『Python ではじめる機械学習：scikit-learn で学ぶ特徴量エンジニアリングと機械学習の基礎』. オライリー・ジャパン）

【研究発表 Session 9　DDL】
【研究発表 1 】

英語の動詞 - 名詞コロケーション学習に対する DDL の効果

佐竹　由帆（青山学院大学）

コロケーションの知識は重要だが，日本の英語学習者の語彙学習は単語単位になりがちであり，コロケーション指導・学習は十分に行われていない。コーパスを参照して学習するデータ駆動型学習（DDL）の語彙学習に対する効果は様々に検証されているため，本研究は DDL の動詞 - 名詞コロケーション学習に対する効果を検証した。被験者は約 20 名の日本の大学 2 年生で英語中級学習者であり，筆者が選んだ２つの動詞 - 名詞コロケーションを現代アメリカ英語コーパス（COCA）で検索して用例を見る学習を毎週 1 度 10 週間行った。事前事後のテスト結果は，ウィルコクソンの符号順位検定で有意差有り，効果量大で（z=3.63, p=.000, r=.59），コロケーション学習に対する DDL の有効性が示唆された。

主要参考文献

Shin, D. & Nation, P. (2008) Beyond single words: the most frequent collocations in spoken

English, *ELT journal* Vol. 62(4), pp. 339-348.

【研究発表２】

中学生のための Web 版 DDL 支援ツールの開発と活用

西垣知佳子（千葉大学）
赤瀬川史朗（Lago 言語研究所）
川名　隆行（千葉大学教育学部附属中学校）
中井　康平（千葉大学教育学部附属中学校）
見目　慎也（千葉大学教育学部附属中学校）
山﨑　達也（千葉大学教育学部附属中学校）

DDL は大学生を対象とする活用事例は多いが，中学生のような入門期では世界的に少ない。その理由には，入門期レベルに適したコーパスと使いやすい検索ソフトの不足がある。

そこで発表者らは，レベルに配慮した teaching-oriented corpus を作成し，その検索と学習を助ける Web 版 DDL 支援ツールを公開した。本研究の目的は，本ツールを使って中学生がどのように英語の学びを深めるか調査し，有効性を確認することであった。国立大学附属中学校の生徒に対して，誤り訂正タスクと Web 版 DDL ツールを組み合わせた授業を実施した。ワークシートを分析した結果，DDL が教師の期待する文法や語彙の気づきを生徒より引き出したことから，学習に有効であると確認した。

主要参考文献

Crosthwaite, P. (2020). *Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-tertiary Learners*. Routledge.

堀正広，赤野一郎.（2015）.（監修），投野由紀夫（編）.『英語コーパス研究シリーズ（第 2 巻）：コーパスと英語教育』ひつじ書房.

Timmis, I. (2015). *Corpus Linguistics for ELT: Research and Practice*. Routledge.

【研究発表３】

Introducing AntConc 4.00: A Fast, Powerful, and Easy-to-Use Corpus Analysis Tool for Small and Large-Scale Corpus Analysis and Data-Driven Learning

Laurence ANTHONY（Waseda University）

AntConc is a corpus analysis tool that has been repeatedly cited to be the most widely

used desktop corpus tool in the world (e.g., Tribble, 2012). It has been downloaded over 2.5 million times by users in over 140 countries and its tutorial videos have been viewed over 500,000 times. While AntConc is a relatively powerful and easy-to-use tool, two of its most commonly cited weaknesses are its speed and handling of medium to large corpora of 10 million words or more. To address these issues, AntConc 4.00 has been completely rewritten on top of a modern database indexing system that scales to corpora of 100 million words and more and allows for results from large corpora to be returned in fractions of a second. Also, the interface has been redesigned to allow for pagination or thinning of large sets of results that are commonly generated with large corpora. In addition, AntConc 4.00 introduces a completely original Key-Word-In-Context (KWIC) concordance view that dramatically simplifies the use and interpretation of this tool. It is anticipated that this new view will become a standard in the field and greatly improve the utility of KWIC concordancing as part of Data-Driven Learning (DDL) approaches.

References

Anthony, L. (2021). AntConc (Version 4.0.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.html.

Tribble, C. (2012). Teaching and language corpora: Quo Vadis? 10th Teaching and Language Corpora Conference (TALC). Warsaw, 11th-14th July 2012.

『英語コーパス研究』投稿規定

（2019年10月改定）

1. 投稿資格

投稿は会員に限る。共著の場合，第一著者は会員であることとし，その他の共著者については会員でなくてもよい。

2. 原稿の種類と長さ

【研究論文】

英文　A4サイズ　1ページあたり35行（文字数の指定はしない），周囲の余白1インチ（25.4mm），17ページ以内（Times New Roman 10.5ポイント使用）

和文　A4サイズ　1ページあたり35行（文字数の指定はしない），投稿時17枚以内（明朝体フォント（游明朝･ヒラギノ明朝など）10.5ポイント使用）

※和文中の英文のフォントについては Times New Roman を原則とする。Century は用いてはならない。

（いずれも Abstract（英文300語以内），図表，注，参考文献目録，付録，謝辞，著者情報などを含む。）

【研究ノート，総説論文・書評論文（Review article, Book review）】

・研究ノート：論文のカテゴリーに属さない小論文や萌芽的な研究，新しい研究開発の成果などをまとめたもの

・総説論文：体系的かつ網羅的に先行研究をまとめたもの

・書評論文：専門書の研究分野への貢献と課題点を明確にしたもの

英文　A4サイズ　1ページあたり35行（文字数の指定はしない），周囲の余白1インチ（25.4mm），12ページ以内（Times New Roman 10.5ポイント使用）

和文　A4サイズ　1ページあたり35行，投稿時12枚以内（明朝体フォント（游明朝･ヒラギノ明朝など）10.5ポイント使用）

※和文中の英文のフォントについては Times New Roman を原則とする。Century は用いてはならない。

（いずれも Abstract（英文300語以内），図表，注，参考文献目録，付録，謝辞，著者情報などを含む。）

【その他（ソフトウェアレビュー，書評（図書紹介），コーパス紹介など）】

研究論文の半分以内の分量

3. 原稿作成時の注意

下記のように投稿者を特定できるような情報，その他，本人の同定につながると考えられる情報は，採用決定後の最終原稿に追記するものとし，投稿時には記載しないこと。

（1）謝辞など
（2）「本論は，英語コーパス学会第X回大会において口頭発表した内容に加筆修正を施したものである。」などの文言
（3）「筆者が収集し，WWW（http://…）で公開しているデータ…」など，筆者情報につながるURL情報など
（4）「拙論（2006）で論じたように…」などと記して，参考文献目録で当該文献を参照している場合，「拙論」ではなく著者（2006）として表記すること。

4. 提出方法など

（1）下記の（A）原稿ファイル（Microsoft Wordで作成したファイルとそのPDFファイル），（B）著者情報ファイル，（C）論文投稿チェックシートの3種類のファイルを電子メール添付で提出。（B），（C）についてはWeb掲載のフォーマットを使用のこと。
（2）電子メールの件名（Subject）は「『英語コーパス研究』投稿原稿（著者氏名）」とすること。
（3）提出先，締め切り期日等に関しては学会Webサイトを参照のこと。

（A）原稿ファイル
a. 提出するファイル名は「原稿題名（著者氏名）」とすること。
b. 原稿題名の前に「論文」，「研究ノート」，「総説論文」，「書評論文」，「コーパス紹介」などの種類を明記すること。
c. 原稿本体の冒頭には上記種類の別と題名のみを記すこと。

（B）著者情報ファイル：「著者情報（著者氏名）」
a. 和文原稿の場合は英文タイトル，英文原稿には和文タイトル
b. 著者氏名（ふりがな・ローマ字表記）
c. 所属
d. 郵便番号・住所・電話番号
e. 電子メールアドレス

（C）論文投稿チェックシート：「論文投稿チェックシート（著者氏名）」
Web掲載のチェックシートの必要項目すべてに☑を入れること。

5. スタイル

投稿論文は，研究論文，研究ノート，総説論文・書評論文の別，また，和文・英文の別にかかわらず,『英語コーパス研究』スタイルシートに従い執筆することとする。

6. 掲載論文等の電子化

掲載された論文等の著作者は，論文等を電子化して学会ホームページで公開することに同意する。

7. 著作権

掲載された論文等の著作権は,本学会に帰属する。本学会は掲載論文等を印刷媒体・電子媒体で公開する権利を有するものとする。ただし，著作者が自著論文等を自分のホームページに掲載したり，自著の本に転載したりすることは妨げない。

8. 研究倫理

投稿にあたっては，下記文書などを参照し，不正行為のないようにすること。
独立行政法人科学技術振興機構『研究者のみなさまへ～研究活動における不正行為の防止について～』https://www.jst.go.jp/contract/kisoken/h25/others/h25s805others131120.pdf

英語コーパス学会会則

（名称）

第１条　本会は「英語コーパス学会」（Japan Association for English Corpus Studies，略称 JAECS）と称する。

（目的）

第２条　本会は英語コーパス及びコーパスツールの開発・評価・利用に関わる研究，また，英語コーパスを用いた言語研究・言語教育研究・関連研究を促進することを目的とする。

（事業）

第３条　本会は前条の目的を達成するために，次の事業を行う。

(1) 大会・研究会等の開催
(2) 学会誌・学会報等の発行
(3) その他本会の趣旨に沿う事業

（会員）

第４条　本会の会員は一般会員，学生会員，団体会員，賛助会員，功労会員及び名誉会員よりなる。

(1) 一般会員は本会の趣旨に賛同する個人とする。
(2) 学生会員は本会の趣旨に賛同する個人のうち，大学又は大学院に籍を置く学生とする。
(3) 団体会員は本会の趣旨に賛同する大学，研究所，図書館その他の研究・教育団体とする。
(4) 賛助会員は本会の趣旨に賛同する企業等とする。
(5) 功労会員は本会の活動に長く寄与した個人とする。功労会員の規程は別に定める。
(6) 名誉会員は本会の活動に特別に寄与した個人とする。

（会費）

第５条　本会の会費について以下の通り定める。

(1) 会員は所定の会費を納めるものとする。
(2) 会費の額については次の通りとする。
　一般会員　年額　5,000円（在外会員は年額12,000円）
　学生会員　年額　2,000円（在外会員は年額10,000円）
　団体会員　年額　5,000円
　賛助会員　年額　15,000円
(3) 会費は入会時点又は会計年度開始時点で納入する。
(4) ２年間にわたって会費納入がない場合は会員の資格を失う。

(5) 名誉会員，功労会員，顧問からは会費を徴収しない。

(会計年度)
第６条　本会の会計年度は４月１日に始まり，翌年３月31日をもって終わる。

(組織)
第７条　本会に執行部，事務局，役員会，学会誌編集委員会，学会賞選考委員会，大会実行委員会，研究会（SIG）を置く。
(1) 執行部は会長，副会長，事務局長，事務局員で構成し，本会全体にかかわる事業を執行・監督する。
(2) 事務局は事務局長及び事務局員で構成し，本会の事務を執行する。
(3) 役員会は役員で構成し，本会にかかる諸問題を審議・決定する。
(4) 学会誌編集委員会は学会誌の刊行にかかる業務を担当する。学会誌編集委員会の規程は別に定める。
(5) 学会賞選考委員会は学会賞・奨励賞の選考にかかる業務を担当する。学会賞選考委員会の規程は別に定める。
(6) 大会実行委員会は大会の企画・準備・実施にかかる業務を担当する。大会実行委員会の規程は別に定める。
(7) 研究会（SIG）は会員のうち，希望する者によって構成し，それぞれが掲げる研究目的に応じた活動を行う。研究会の規程は別に定める。

(役員)
第８条　本会に次の役員をおく。
(1) 会　長　１名
(2) 副会長　若干名
(3) 理　事　若干名
(4) 幹　事　若干名

(役員の任期・定年)
第９条　役員の任期は以下の通りとする。
(1) 会長・副会長の任期は２年とし，引き続き２期までの再任を妨げない。
(2) 理事・幹事の任期は２年とし，再任を妨げない。
(3) 任期は当該年度の４月１日から起算する。
(4) 役員の定年を70歳とする。任期の途中で定年に達したときは当該年度の終了まで，その任にあたる。

(役員の任務)
第10条　役員の任務は以下の通りとする。
(1) 会長は本会を代表し，会務を統括する。会長は総会・役員会を招集し，これを主宰する。

(2) 副会長は会長の命ずる職務を所掌するとともに，会長を補佐し，必要に応じて会長の職務を代行する。
(3) 理事は役員会に出席し，本会の運営に関わる重要事項を審議･議決する。
(4) 幹事は役員会に出席し，理事を補佐し，本会の運営に関わる重要事項を審議・議決する。

(役員の選出)
第11条　役員は役員会における投票によって決定する。

(役職員)
第12条　本会に次の役職員をおく。

(1) 顧問	若干名
(2) 事務局長	1名
(3) 事務局員	若干名
(4) 監査	1名
(5) 学会誌編集委員会委員長	1名
(6) 学会誌編集委員	若干名
(7) 学会賞選考委員会委員長	1名
(8) 学会賞選考委員	若干名
(9) 大会実行委員会委員長	1名
(10)大会実行委員	若干名

(役職員の任期・定年)
第13条　役職員の任期は以下の通りとする。
(1) 顧問の任期は終身とする。
(2) 事務局長・事務局員，監査，学会誌編集委員会委員長及び委員，学会賞選考委員会委員長及び委員の任期は2年とし，引き続き2期までの再任を妨げない。任期は当該年度の4月1日から起算する。
(3) 大会実行委員会委員長及び委員の任期は，役員会で承認された日から当該大会に関係する業務の終了時までとする。
(4) 顧問を除く役職員の定年を70歳とする。任期の途中で定年に達したときは当該年度の終了まで，その任にあたる。

(役職員の任務)
第14条　役職員の任務は以下の通りとする。
(1) 顧問は役員会の求めに応じて学会運営への助言を行う。
(2) 事務局長は事務局を主宰し，本会の事務を執行・監督する。
(3) 事務局員は事務局長の指示の下，必要な業務を執行する。
(4) 監査は本会の会計及び運営が適切になされているか精査し，その結果を総会で報告する。

(5) 学会誌編集委員会委員長は学会誌編集委員会を主宰し，学会誌の刊行にかかる業務を執行・監督する。
(6) 学会誌編集委員は委員長の指示の下，必要な業務を執行する。
(7) 学会賞選考委員会委員長は学会賞選考委員会を主宰し，学会賞・奨励賞の選考にかかる業務を執行・監督する。
(8) 学会賞選考委員は委員長の指示の下，必要な業務を執行する。
(9) 大会実行委員会委員長は大会実行委員会を主宰し，大会の企画・準備・実施にかかる業務を執行・監督する。
(10) 大会実行委員は委員長の指示の下，必要な業務を執行する。

(役職員の選出)
第15条　役職員は会長が推薦し，役員会で承認する。役職員と役員の兼務を妨げない。

(会議)
第16条　本会は以下の会議を開催する。
(1) 総会は会長の招集により，原則として年１回以上開催し，会則の改定，予算・決算その他重要事項を審議する。なお，電子メールやその他の手段を用いた総会の開催も可能とする。総会での議決は出席者の過半数による。
(2) 役員会は会長の招集により，原則として年２回以上開催し，本会の運営にかかる諸問題を審議し，決定する。なお，電子メールやその他の手段を用いた役員会の開催も可能とする。役員会での議決は出席者の過半数による。
(3) 事務局会議は事務局長の判断の下，不定期に開催する。
(4) 学会誌編集委員会，学会賞選考委員会，大会実行委員会は各委員長の判断の下，不定期に開催する。

付則
(1) 本会則は2020年４月１日から施行する。
(2) 本会則は2021年４月１日から改正施行する。

(備考)
(1) 本会は1993年４月１日に「英語コーパス研究会」として発足し，1997年４月１日に「英語コーパス学会」に改組されて現在に至る。
(2) 本会の事務局を岐阜市立女子短期大学英語英文学科小島ますみ研究室(〒501-0192 岐阜市一日市場北町７番１号）に置く。

英語コーパス研究（第29号）
【2022年5月31日発行】

編集・発行　ⓒ2022 英語コーパス学会
〒501-0192 岐阜市一日市場北町 7 番 1 号
岐阜市立女子短期大学英語英文学科　小島ますみ研究室気付
E-mail（事務局長）: jaecs.hq@gmail.com
Twitter: @JAECS2012
Website: http://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.29
2022

JAPAN ASSOCIATION FOR ENGLISH CORPUS STUDIES