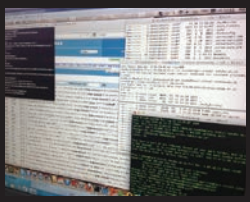
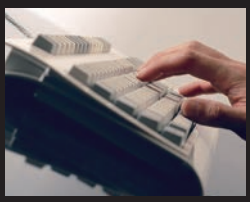
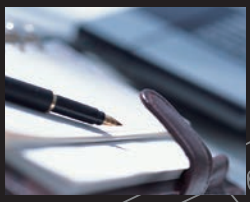
ISSN 1340-301 X

JAECS
Japan Association for English Corpus Studies

2018
英語コーパス学会

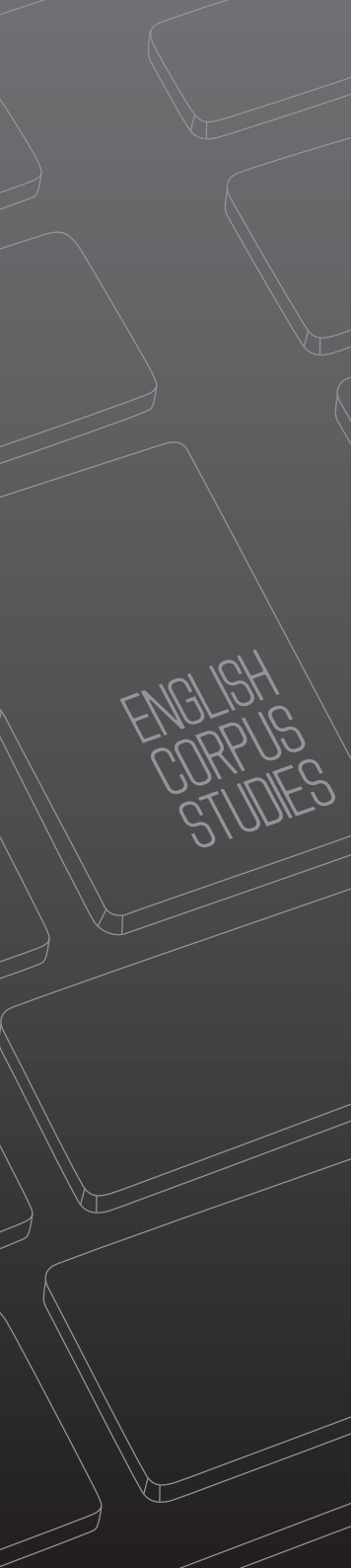






英語コーパス研究
第25号

English Corpus Studies 25



2017年度
英語コーパス学会役員

会　　長：投野由紀夫

副 会 長：井上　永幸

事務局長：石井　康毅

会　　計：宇佐美裕子，小島ますみ

顧　　問：赤野　一郎，中村　純作

理　　事：新井　洋一，家入　葉子，石井　康毅，石川慎一郎，
石川　保茂，井上　永幸，梅咲　敦子，岡田　　毅，
地村　彰之，園田　勝英，高橋　　薫，滝沢　直宏，
田畑　智司，中條　清美，塚本　　聡，投野由紀夫，
西村　秀夫

監　　事：加野まきみ

事務局補佐：内田　　諭，大谷　直輝

学会賞選考委員会：
＜委 員 長＞西村　秀夫
＜委　　員＞新井　洋一，地村　彰之，高橋　　薫，田畑　智司，
中條　　清美

大会企画委員会：
＜委 員 長＞金澤　俊吾
＜委　　員＞石川　有香，小島ますみ，滝沢　直宏，長　加奈子，
西村　秀夫，藤原　康弘

編集委員会：
＜委 員 長＞中尾　佳行
＜委　　員＞家入　葉子，瀬良　晴子，田畑　智司，能登原祥之，
水野　和穂，水本　　篤

論文査読委員：
家入　葉子，瀬良　晴子，田畑　智司，能登原祥之，
水野　和穂，水本　　篤，石川　保茂，井上　永幸，
梅咲　敦子，滝沢　直宏，地村　彰之，中條　清美，
西村　英夫

ISSN 1340-301 X

英語コーパス研究

第 25 号

英語コーパス学会

2018

目　次

論文

研究ノート

特別講演

シンポジウム

資料

「論文」

Conceptual Metaphors and Metonymies of Near-Synonyms of ANGER

Yuki MINAMISAWA

Abstract

This article analyzes conceptual metaphors and metonymies involving near-synonyms of ANGER: *anger* and *rage*. Within conceptual metaphor theory, many studies have been conducted to describe metaphors/metonymies conceptualizing the emotions, and ANGER has attracted much attention in previous studies. A certain number of metaphors/metonymies have been proposed for the emotion, and it is generally agreed that the central metaphor for ANGER is ANGER IS A HOT FLUID IN A CONTAINER. However, few studies have focused on differences between near-synonyms, partly because of the assumption that the words representing similar (or the same) emotions have similar metaphors. Therefore, this study aims to examine the differences in metaphors/metonymies of near-synonyms from the perspective of their centrality to the emotion. The present study uses MI-scores to determine the centrality. The data extracted through the MI-scores shows that there are differences between *anger* and *rage* in the metaphors with which they are strongly associated. This result is further supported by an examination of their metonymical collocates.

1. Introduction

This study investigates figurative expressions of ANGER within conceptual metaphor theory (Lakoff and Johnson, 1980, henceforth CMT). In this theory, metaphor is regarded not as a matter of language, but as a matter of thought and action. We understand abstract concepts, which are often invisible and intangible, in terms of other more concrete concepts.[1] Emotion is one such abstract concept, and many studies have been conducted to elucidate what metaphors/metonymies are used to conceptualize emotions. Among the emotions, ANGER has attracted much attention, and a certain

number of metaphors and metonymies have been proposed in studies (e.g., Kövecses, 1990, 2000; Lakoff, 1987).

In previous studies, however, little attention was paid to the differences between near-synonyms, although there are several different words representing ANGER. For example, previous studies like Lakoff (1987) and Kövecses (1990) did not distinguish between *anger* and *rage*. Although the assumption that the words representing similar (or the same) emotions have similar metaphors and metonymies might be at play, little discussion has been devoted to verifying its validity. Therefore, the present study aims to ascertain whether near-synonyms are associated with the same conceptual metaphors and metonymies.

2. Previous Studies

2.1 ANGER Metaphors and Metonymies

Within CMT, much attention has been paid to emotions, and it was found that metaphors/metonymies play a crucial role in conceptualizing them. For example, ANGER is conceptualized in terms of various concepts like A HOT FLUID IN A CONTAINER (henceforth the FLUID metaphor), FIRE (the FIRE metaphor), A DANGEROUS ANIMAL (the ANIMAL metaphor), and A NATURAL FORCE (the NATURE metaphor). Below are instantiations of each metaphor (1a-d).

(1) a. She is *boiling with* anger.
b. His anger is *smoldering*.
c. He *unleashed* his anger.
d. It was a *stormy* meeting. (Kövecses, 2000: 21)

Furthermore, such metaphors are often based on cultural models of physiological and behavioral responses. For example, ANGER is often regarded as something hot (1a,b). This is ANGER IS HEAT, the most general metaphor for ANGER, and it is based on human physiological effects such as BODY HEAT (2a) and AGITATION (2b).

(2) a. Don’t get *hot under the collar*.

b. She was *shaking with anger*. (Kövecses, 1990: 52)

These physiological and behavioral responses provide the basis of conceptual metaphors, and also metonymically indicate the emotion.

Moreover, such metaphors and metonymies converge on a certain prototypical cognitive model of the emotion (Lakoff, 1987; Kövecses, 1990, 2000). The model has a temporal dimension and consists of different stages. Kövecses (1990) maintains that the prototype of the emotion involves the following temporally and causally connected stages: 1. Cause, 2. Emotion exists, 3. Attempt at control, 4. Loss of control, and 5. Action (+ 0. Emotional calmness).

2.2 Main Issues of Emotion Metaphors

With regards to research on emotion metaphors/metonymies, the following two issues have been discussed: Which conceptual metaphors/metonymies are mainly used for a particular emotion (Issue 1), and which metaphors are the most central to the emotion (Issue 2). In the case of ANGER, there has been much discussion of these issues. In relation to Issue 1, Kövecses (2000: 21) gives 12 main metaphors, including the afore-mentioned metaphors. However, it is not absolutely clear why these 12 metaphors are considered the main metaphors. In fact, Stefanowitsch (2006) found several metaphors through his corpus-based method that were not mentioned in previous studies.

Concerning Issue 2, it is generally agreed that the central metaphor for ANGER is the FLUID metaphor. As reasons for this, Kövecses (1990) states that a variety of words and expressions belong to this metaphor, and that the metaphor productively carries over knowledge from the source (A HOT FLUID IN A CONTAINER) to the target (ANGER).

(3) a. His pent-up anger *welled up* inside her.
b. I *suppressed* my anger.
c. When I told him, he just *exploded*. (Kövecses, 1990: 54, 55)

The expression in (3a) indicates an increase in the intensity of ANGER by the rise of a fluid, while (3b) highlights the control aspect by comparing it to keeping back a fluid in

a container. When the pressure becomes too high, the container explodes (3c), and this is used for loss of control. In this way, the FLUID metaphor captures different aspects of ANGER, and is therefore regarded as the central metaphor for ANGER.

2.3 Near-Synonyms

In recent years, an increasing number of studies of emotion metaphors have been performed on the basis of corpus data, as this provides an empirical basis for studying conceptual metaphors from a linguistic perspective. As Stefanowitsch (2006) shows, we might thereby find some metaphors that are not mentioned in previous studies. Moreover, frequency data provides valuable insight into determining the importance of conceptual metaphors (Stefanowitsch, 2006).

However, little attention has been paid to the issue of near-synonyms in previous studies. This is partly because of the assumption that the words representing similar (or the same) emotions should have similar metaphors, but also because of the difficulty in comparing near-synonyms only on the basis of introspective data.

Recently, several corpus-based studies have focused on the similarities and differences of metaphors between near-synonyms. For ANGER, Suzuki (2010) demonstrates that different words representing ANGER are connected to different metaphors. For example, he states that the number of metaphorical expressions for *anger* and *fury* motivated by ANGER IS FIRE is high, while there are few expressions featuring *rage* and *wrath* that are motivated by this metaphor. Suzuki describes some differences between near-synonyms of ANGER, although the focus of the discussion is mainly on ANGER IS HEAT/FIRE and other metaphors like the ANIMAL metaphor are not much discussed.

In this respect, Turkkila (2014) discusses metaphors more inclusively based on data from the Corpus of Contemporary American English. Turkkila's study examines the assumption that near-synonyms occur with the same metaphors. Turkkila provides 37 conceptual metaphors (pp. 136−137, Table 2) and, according to the study, the near-synonyms *anger*, *rage*, and *fury* more or less occur with the same metaphors. Although *wrath* is different in lacking six of these metaphors, such as ANGER IS A DISEASE and ANGER IS A PLANT, Turkkila implies that this is because *wrath* is by far the least frequent in the corpus and is a formal word.

While Turkkila's study shows that *anger*, *rage*, and *fury* are almost the same in

terms of the metaphors applicable to them (Issue 1), there is room for discussion as to which metaphors are more central (Issue 2). According to the list provided in the study (pp. 136−137), the four most frequent source domains for ANGER are A POSSESSION (*have anger*), A PLACE (*in anger*), A MOVING OBJECT (*anger toward X*), and AN OBJECT (*anger against X*). The fifth most frequent is FIRE (*anger burn inside X*), and AN OBJECT IN A LOCATION (*anger in X*) is the sixth. Instances that feature the mappings ANGER IS A POSSESSION/A PLACE/A MOVING OBJECT/AN OBJECT/AN OBJECT IN A LOCATION make up 70.01% of all metaphorical mappings representing ANGER in the corpus.[2] Such results are a little problematic since these metaphors are considered highly general. The words used in these metaphorical expressions are very frequently used by themselves, and these metaphors are shared not only by near-synonyms of ANGER, but also by other emotions and by other abstract concepts. On this matter, Kövecses (2011) emphasizes that the metaphors that contribute to a greater extent to the structure of abstract concepts are specific ones like the FLUID metaphor, and contends that the quantitative advantage of the corpus-based method does not necessarily lead to a qualitative advantage.

3. Materials and Methods

3.1 Metaphorical Pattern Analysis

Although studies of conceptual metaphors are now increasingly conducted using corpora, corpus-based methods do not seem appropriate for analyzing emotion metaphors. This is because we must choose a particular word form when searching metaphorical expressions, although conceptual mappings are originally not linked to particular linguistic forms (Stefanowitsch, 2006: 64). In other words, we cannot know in advance which words and phrases are used for a particular metaphor.

To cope with this problem, Stefanowitsch (2006) suggests Metaphorical Pattern Analysis (henceforth MPA). According to Stefanowitsch, a metaphorical pattern is "a multi-word expression from a given source domain (SD) into which one or more specific lexical item(s) from a given target domain (TD) have been inserted" (p. 66).

(4) a. He was *bursting with* anger.
b. I was *fuming*. (Kövecses, 1990: 54)

Following the above definition, (4a) is a metaphorical pattern since the example includes both a source-domain word (*bursting*) and a target-domain word (*anger*). On the other hand, (4b) is not a metaphorical pattern, because the sentence does not contain a target-domain word (an emotion word). Stefanowitsch claims that investigating metaphorical patterns like (4a) makes it possible to perform target-domain-oriented studies based on corpus data. Although the metaphorical expressions without a target-domain word as in (4b) are not taken into account, the data Stefanowitsch provides show that with MPA it is possible to find examples of almost all of the metaphors identified by the previous introspective methods, as well as other metaphors.

Therefore, this study was conducted on the basis of metaphorical patterns, and also attempts to analyze metonymies using the same MPA method. In extracting metonymical expressions from a corpus, however, there is another problem.

(5) a. She was *quivering* with rage.
b. Don't get *hot under the collar*. (Kövecses, 1990: 52)

The expression in (5a) is an instantiation of the metonymy AGITATION FOR ANGER, while (5b) is yielded by BODY HEAT FOR ANGER. Following the above method, it is possible to retrieve expressions like (5a) since they include an emotion word, but expressions like (5b) cannot be retrieved because of the absence of an emotion word. However, since the expressions of the former type include an emotion word, they are not genuinely metonymical (rather literal) as Oster (2010) admits of such expressions. Nevertheless, Oster claims that these expressions provide us with an insight into which physiological and behavioral responses are prevalent in the conceptualization of these emotions. In fact, many previous studies (e.g., Lakoff, 1987; Kövecses, 1990, 2000) regard such expressions as instantiations of metonymies. In light of this, expressions like (5a) may be regarded as instantiations of metonymies to the extent that the physiological and behavioral responses are strongly connected to the emotion.

3.2 Metaphorical Pattern as Collocation

In brief, the analysis in this study is based on metaphorical/metonymical expressions that include an emotion word as in (4a) and (5a). In these expressions, an

emotion word and a metaphorically/metonymically used word always co-occur. In (4a), *anger* co-occurs with the metaphorical collocate *bursting*, and in (5a), *rage* co-occurs with the metonymical collocate *quivering*. Since collocation is "the occurrence of two or more words within a short space of each other in a text" (Sinclair, 1991: 170), such a metaphorical/metonymical expression is a specific type of collocation.

Considering that collocation is often measured using statistical methods (Hunston, 2002), it is expected that such figurative expressions can be measured in the same way. Although there are different measurements for calculating collocation, this study uses the Mutual Information score[3] (henceforth MI-score). The purpose of this study is to analyze metaphors/metonymies representing ANGER, as they play an essential role in how we understand the concept. Since the MI-score is appropriate for investigating collocations that are semantically connected to each other (Akano, 2009), it is expected that the MI-score effectively retrieves metaphorically/metonymically used collocates.

Using the MI-score, this study attempts to measure the centrality of metaphors and metonymies. In general, the MI-score indicates the strength of a collocation (Hunston, 2002: 71), and this leads to the idea that a collocate with a higher score is considered to be more central, namely, more strongly associated with the emotion (Criterion 1). Furthermore, centrality can be defined from the perspective of the number of significant collocates. According to Hunston (ibid.), co-occurrences may be considered significant when the MI-score is three or greater. By categorizing the significant collocates into various metaphors/metonymies, we can create a list of which metaphors/metonymies contain such significant collocates and then define the metaphors/metonymies that contain more significant collocates as more central to the emotion (Criterion 2).

In the following sections, the centrality of metaphors/metonymies is measured according to the above criteria. Presumably, these criteria provide valuable information for describing the differences between near-synonyms from the perspective of Issue 2.

3.3 Methods

In this research, the analysis is based on the data extracted from the British National Corpus (BNC*web*, CQP-Edition, Version 4.3, henceforth BNC), a balanced corpus of British English containing 100 million words. In extracting metaphorical/

metonymical collocates from the corpus (both written and spoken texts), *anger* and *rage* were selected as search words, for the reason that they are relatively frequent in the corpus and many previous studies have dealt with these two words without distinction in the analysis of ANGER. Furthermore, the present analysis is confined to singular forms in order to exclude, for example, the name of the city in western France, *Angers*. Thus, the search was conducted in the form "anger_NN1" and "rage_NN1". As for the window size (span), 4:4 is used (Krishnamurthy, 2003), and the words extracted are lemmatized. Finally, words which occur only in one text are excluded, since they are likely characteristic of the author.

As described above, the present study adopts the MI-score to determine the centrality of metaphors/metonymies to the emotion, and only takes the significant collocates ($MI \geqq 3$) into account. However, since the MI-score becomes unstable when the number of co-occurrences is very small (Church and Hanks, 1990), the present research does not consider collocates if they occur fewer than four times.

All of the significant figurative collocates extracted in this way are then classified into various metaphors or metonymies. In classifying these collocates, previous studies such as Lakoff (1987), Kövecses (1990, 2000), and Stefanowitsch (2006) are the most often cited. For example, Kövecses regards the expression *boiling with anger* (2000: 21) as an instantiation of the FLUID metaphor, and *shaking with anger* (1990: 52) as that of the metonymy AGITATION FOR ANGER. Accordingly, the present study classifies the verb *boil* into the FLUID metaphor, and *shake* into AGITATION FOR ANGER.

4. Results and Discussion

The above procedure retrieved 150 significant collocates for *anger* and 63 for *rage*. Of these significant collocates, 75 collocates of *anger* and 40 of *rage* are instantiations of metaphors/metonymies. Since over 30 significant collocates of *anger* and about 15 of *rage* are emotion words like *frustrated*, *sorrow*, and *hatred*, which do not directly represent ANGER, it is possible to state that many of the significant collocates are instantiations of metaphors/metonymies. Considering that the MI-score tends to extract collocates that are semantically connected to the search words, it appears that these corpus data provide supporting evidence for the CMT argument that

we understand abstract concepts like emotions with the help of metaphors and metonymies.

4.1 Top Collocates

Table 1 shows the top 20 collocates of *anger* and *rage* according to MI-score.

Table 1: Top 20 collocates of a*nger* and r*age*

No.	*anger*	Freq.	MI	*rage*	Freq.	MI
1	*vent*_(V)	26	9.99	*incandescent*_(A)	7	10.18
2	*sethe*_(V)[4]	9	8.91	*bellow*_(S)	5	9.87
3	*suppressed*_(A)	11	8.81	*vent*_(V)	8	9.60
4	righteous_(A)	11	8.65	*contort*_(V)	5	9.56
5	*seethe*_(V)	4	8.43	*suppressed*_(A)	7	9.47
6	*contort*_(V)	5	8.24	*howl*_(S)	8	9.23
7	*pent-up*_(A)	4	8.14	*impotent*_(A)	7	9.03
8	*livid*_(A)	5	8.08	*speechless*_(A)	5	8.71
9	*well*_(V)	6	7.78	*bristle*_(V)	5	8.59
10	*simmering*_(A)	5	7.75	righteous_(A)	4	8.51
11	frustration_(S)	49	7.65	*murderous*_(A)	5	8.50
12	resentment_(S)	31	7.57	*towering*_(A)	5	8.30
13	placate_(V)	4	7.30	*choke*_(V)	11	7.71
14	*bubble*_(V)	8	7.28	drunken_(A)	7	7.56
15	hurt_(S)	5	7.22	frustration_(S)	16	7.35
16	*outburst*_(S)	11	7.21	humiliation_(S)	6	7.29
17	*subside*_(V)	10	7.15	*roar*_(S)	5	7.17
18	*surge*_(S)	16	7.08	*scarlet*_(A)	5	7.01
19	*speechless*_(A)	4	7.08	*tremble*_(V)	11	6.99
20	abate_(V)	4	6.94	*fit*_(S)	8	6.96

In the table, all of the figurative collocates are highlighted in italics. The symbols (S), (V), and (A) correspond to nouns, verbs, and adjectives. Obviously, many of the top collocates of *anger* and *rage* are instances of metaphors or metonymies (13 collocates for *anger* and 16 for *rage*).

However, there are interesting differences between *anger* and *rage*. Of the top 20 collocates, as many as eight collocates of *anger* are associated with the FLUID metaphor[4] (shaded in light gray), which indicates that *anger* is strongly associated with this metaphor. The result is in line with the observations of most previous studies. However, in the case of *rage*, only two belong to the FLUID metaphor. Instead, five

collocates go into the ANIMAL metaphor (shaded in dark gray). Such collocates do not appear among *anger*'s top 20 collocates.

According to the criterion that a collocate with a higher score is considered to be more strongly associated with the emotion (Criterion 1), *anger* is strongly associated with the FLUID metaphor, whereas *rage* is strongly associated with the ANIMAL metaphor.

4.2 Conceptual Metaphors

When all of the significant metaphorical collocates are categorized under different metaphors, the difference becomes clearer.

Table 2: Metaphorical collocates significantly connected to a*nger* and r*age*

Source	*anger* [57 collocates]	*rage* [27 collocates]
FLUID	*vent*_(V), *seethe*_(V), *suppressed*_(A), *pent-up*_(A), *well*_(V), *simmering*_(A), *bubble*_(V), *outburst*_(S), *evaporate*_(V), *vent*_(S), *seep*_(V), *explode*_(V), *suppress*_(V), *boil*_(V), *burst*_(S), *explosion*_(S), *drain*_(V), *burst*_(V), *inside*_(P), *rise*_(V), *fill*_(V) [21 (36.8%)]	*vent*_(V), *suppressed*_(A), *boil*_(V), *burst*_(S), *explode*_(V), *burst*_(V), *fill*_(V) [7 (25.9%)]
FIRE	*glitter*_(V), *flare*_(V), *flash*_(S), *blaze*_(V), *fuel*_(V), *spark*_(V), *flash*_(V), *flame*_(S), *burn*_(V) [9 (15.8%)]	*incandescent*_(A), *flash*_(S), *consume*_(V) [3 (11.1%)]
ANIMAL	*uncontrollable*_(A), *howl*_(S), *rouse*_(V), *arouse*_(V), *fierce*_(A), *growing*_(A), *control*_(V), *violent*_(A) [8 (14.0%)]	*bellow*_(S), *howl*_(S), *bristle*_(V), *murderous*_(A), *roar*_(S), *roar*_(V) [6 (22.2%)]
NATURE	*subside*_(V), *surge*_(S), *surge*_(V), *wave*_(S) [4 (7.0%)]	*tide*_(S) [1 (3.7%)]
Others	*impotent*_(A), *fit*_(S), *direct*_(V), *mixture*_(S), *melt*_(V), *swallow*_(V), *fade*_(V), *vanish*_(V), *pure*_(A), *convey*_(V), *stir*_(V), *deep*_(A), *widespread*_(A), *cold*_(A), *depth*_(S) [15 (26.3%)]	*impotent*_(A), *fit*_(S), *towering*_(A), *icy*_(A), *blind*_(A), *fly*_(V), *mad*_(A), *beside*_(P), *possess*_(V), *cold*_(A) [10 (37.0%)]

As the table shows, over one-third of *anger*'s metaphorical collocates are associated with the FLUID metaphor (36.8%), followed by the FIRE metaphor (15.8%), the ANIMAL metaphor (14.0%), and the NATURE metaphor (7.0%).

The distribution of the significant metaphorical collocates of *rage* is somewhat different. The table shows that *rage* is strongly associated with the ANIMAL metaphor (22.2%), as well as the FLUID metaphor (25.9%). Most of the collocates of *rage* still go into the FLUID metaphor, but almost as many collocates belong to the ANIMAL metaphor. The result indicates that, although *anger* and *rage* are similar in the types of metaphors they are associated with, they are different in the metaphors with which they are strongly associated. On the basis of the criterion that a metaphor that contains more significant collocates is more central to the emotion (Criterion 2), both *anger* and *rage* are most strongly associated with the FLUID metaphor. Nevertheless, the data clearly shows that *rage* is almost as strongly associated with the ANIMAL metaphor.

4.2.1 ANGER IS A HOT FLUID IN A CONTAINER

Kövecses (2000) defines the FLUID metaphor as the central metaphor for ANGER, and this is confirmed by the present analysis. Each sentence extracted from the BNC is shown with its filename. In addition, the emotion word is underlined and the metaphorical/metonymical collocate (or phrase) is highlighted in italic.

(6) a. Indeed, one of the worst things you can do with anger is *suppress* it.(AYK 642)
b. She could feel the anger *boiling* up inside her. (CH4 265)
c. Then suddenly he seemed to *explode* with anger. (GV7 214)
d. Children give *vent* to their anger in various ways. (B10 1322)
e. To his own surprise all his anger against Edouard had *evaporated*. (C8X 1613)

A variety of collocates of the FLUID metaphor can be found through the MI-score method, and they represent different aspects of ANGER. For instance, (6a) implies an angry person trying to keep his or her anger back, while (6b) indicates very intense anger. Once anger becomes too intense and cannot be held back any more, the person loses control, as in (6c). Before losing control, an angry person can let anger out like in (6d), and anger sometimes disappears as in (6e).

In the case of *rage*, however, the significant collocates represent rather specific aspects.

(7) a. 'I'm not at all impressed, Maggie,' he rasped, *filled* with unusual rage. (HGK 2661)

b. Lewis was *boiling* with rage and misery and shock. (CDB 746)
c. Her rage *burst* over him like hailstones. (HH9 2339)

Many of *rage*'s collocates represent either intense anger (7a,b) or loss of control (7c). There are no significant collocates corresponding to the disappearance of the emotion. Considering that collocates such as *evaporate* and *drain* do not co-occur with *rage*, we can deduce that the emotion represented by *rage* seems to be so intense that it cannot disappear without an act of retribution.

4.2.2 ANGER IS A DANGEROUS ANIMAL

In Lakoff (1987), ANGER IS A DANGEROUS ANIMAL is one of the main metaphors for ANGER. Table 2 shows that both *anger* and *rage* have a strong association with this metaphor.

The significant collocates of *anger* for this metaphor represent various aspects of the emotion.

(8) a. They *aroused* anger and she felt uncomfortable with it, shifting, frowning. (JYD 305)
b. All he could do now was keep steady despite his *growing* anger. (H86 1622)
c. Immediately the anger and irritation he had brought with him from the house erupted in a *howl* of anger. (FU8 2408)

This metaphor focuses on control of the emotion and the danger to others (Kövecses, 1990). The example in (8a) describes anger being brought into existence and approaching the limit, while (8b) describes anger that is growing more intense. In (8c), the angry person behaves in an angry way. For this aspect, Kövecses proposes ANGRY BEHAVIOR IS AGGRESSIVE ANIMAL BEHAVIOR, which is an extension of the ANIMAL metaphor.

Interestingly, *rage* is more strongly associated with the ANIMAL metaphor than *anger*, and most of its significant collocates represent an angry behavior.

(9) a. A great *bellow* of inhuman rage froze his hand in mid air. (BPA 2920)
b. George felt the rage *roar* in his head. (FAB 1243)

c. Sir John pushed back his chair, his red face *bristling* with rage. (H90 1103)
d. Mandeville looked down, his eyes glowing with a *murderous* rage. (H90 1733)

Within the results, many of the significant collocates represent the loud cry of an animal (9a,b). In (9a), *rage* collocates with the adjective *inhuman*, which emphasizes the characteristics of *rage* as a dangerous animal. The examples in (9c,d) describe anger that is very intense. In particular, *murderous* indicates that the anger is very dangerous. Considering that *rage* co-occurs with *murderous* five times while *anger* collocates with it only once, we can say that the emotion represented by *rage* is very intense and can be dangerous to others. For the verb *bristle*, Kövecses (1990: 63) gives the example: "*he was bristling with anger*." However, *bristle* is not significantly connected to *anger* but only to *rage*.

Another interesting point regarding the ANIMAL metaphor concerns body-part nouns. Importantly, *anger* significantly collocates with *face* (MI=3.91) and *eye* (MI=3.26), and *rage* with *tooth* (MI=4.24) and *face* (MI=3.84). Here, it is worth noting that *tooth* is strongly connected to *rage*.

(10) a. Kate *ground her teeth* in helpless rage. (HGM 915)
b. 'I expect he'd *gnash his teeth* in impotent rage,' said Beuno. (G0X 2044)

These expressions are similar to an example Kövecses (1990: 63) gives: "*he began to bare his teeth*." Kövecses categorizes this expression as ANGRY BEHAVIOR IS AGGRESSIVE ANIMAL BEHAVIOR. Seemingly, expressions such as *grind one's teeth* (10a) and *gnash one's teeth* (10b) fall under the same metaphor, although these expressions are metonymical when the angry person literally grinds his or her teeth. The collocate *tooth* is not metaphorical by itself, but expressions including this collocate can be metaphorical, and they surely evoke a dangerous animal. The significant body-part collocates also indicate that *rage* is highly associated with the ANIMAL metaphor.

4.2.3 ANGER IS FIRE

With regard to the FIRE metaphor, both *anger* and *rage* have several significant collocates (11a-d).

(11) a. But as she looked at him, a tiny spark of anger *flared* within her. (JY5 836)
b. Their anger has been *fuelled* by plans to build a THIRD giant store on their doorstep, which they say would threaten the very fabric of their town. (K1U 1351)
c. He was violently interrupted by a Sally-Anne almost *incandescent* with rage. (HGE 545)
d. Rage *consumed* him. (CJJ 932)

In the FIRE metaphor, the size of the fire corresponds to the intensity of the emotion. The example in (11a) portrays anger that is not intense at first (*a tiny spark*), but then becomes very intense. In (11b), the anger is quite intense from the beginning, and grows even more intense. These expressions indicate that *anger* is used to represent varying degrees of the emotion, such as mild anger, increasing anger, and intense anger, while *rage* generally represents very intense emotion.

4.2.4 ANGER IS A NATURAL FORCE

The above claim is further confirmed by the NATURE metaphor. Below are some examples of this metaphor (12a-c).

(12) a. She felt a sudden *surge* of anger. (JXY 209)
b. Then as soon as it had come, his anger *subsided* and he smiled. (FU8 2049)
c. 'And let me tell you,' she swept on, powered by the hot *tide* of rage flowing through her veins. (JXX 1036)

According to Kövecses (1990), the main focus of this metaphor is lack of personal control over the emotion. This aspect can be seen in both *anger* (12a) and *rage* (12c). Interestingly, the verb *subside*, which corresponds to a decrease in emotional intensity, is also a significant collocate of *anger* (12b). No words representing this aspect can be seen in the list of *rage*'s collocates.

4.2.5 The Prototype Scenario

While *anger* represents different aspects of the emotion, *rage* exclusively represents very intense, violent emotion. In Table 3, all of the significant metaphorical collocates are categorized according to different stages of the emotion. The stages

generally correspond to the prototypical cognitive model of ANGER: 1. Emergence, 2. (Intense) Anger, 3. Attempt at control, 4. Loss of control, 5. Act of retribution, and 6 (0). Disappearance (see Section 2.1).

Table 3: Distribution of the significant metaphorical collocates

Stage	*anger* (57 collocates)	*rage* (27 collocates)
1	*flash*_(S), *spark*_(V), *flash*_(V), *rouse*_(V), *arouse*_(V) [5 (8.8%)]	*flash*_(S) [1 (3.7%)]
2	*seethe*_(V), *pent-up*_(A), *well*_(V), *simmering*_(A), *bubble*_(V), *boil*_(V), *rise*_(V), *fill*_(V), *glitter*_(V), *flare*_(V), *blaze*_(V), *fuel*_(V), *flame*_(S), *burn*_(V), *fierce*_(A), *growing*_(A), *violent*_(A), *surge*_(S), *surge*_(V), *wave*_(S), *convey*_(V), *stir*_(V), *deep*_(A), *depth*_(S) [24 (42.1%)]	*boil*_(V), *fill*_(V), *incandescent*_(A), *consume*_(V), *murderous*_(A), *tide*_(S), *towering*_(A) [7 (25.9%)]
3	*suppressed*_(A), *suppress*_(V), *control*_(V), *swallow*_(V) [4 (7.0%)]	*suppressed*_(A) [1 (3.7%)]
4	*outburst*_(S), *explode*_(V), *burst*_(S), *explosion*_(S), *burst*_(V), *uncontrollable*_(A) [6 (10.5%)]	*burst*_(S), *explode*_(V), *burst*_(V), *blind*_(A), *mad*_(A), *beside*_(P), *possess*_(V) [7 (25.9%)]
5	*vent*_(V), *vent*_(S), *howl*_(S), *fit*_(S) [4 (7.0%)]	*vent*_(V), *bellow*_(S), *howl*_(S), *bristle*_(V), *roar*_(S), *roar*_(V), *fit*_(S) [7 (25.9%)]
6	*evaporate*_(V), *drain*_(V), *melt*_(V), *subside*_(V), *fade*_(V), *vanish*_(V) [6 (10.5%)]	[0]
Others	*seep*_(V), *inside*_(P), *impotent*_(A), *direct*_(V), *mixture*_(S), *pure*_(A), *widespread*_(A), *cold*_(A) [8 (14.0%)]	*impotent*_(A), *icy*_(A), *fly*_(V), *cold*_(A) [4 (14.8%)]

Table 3 shows that various metaphorical expressions are used to represent the different aspects of *anger*. As described in 4.2.1, the FLUID metaphor covers most stages of the emotion. On the other hand, most of *rage*'s significant collocates concentrate on the stages of intense anger (Stage 2), loss of control (Stage 4), and act of retribution (Stage 5). There are no significant collocates corresponding to its disappearance (Stage 6). In addition, few significant collocates correspond to the aspects of emergence (Stage 1) or attempt at control (Stage 3). This indicates again that *rage* represents an intense, violent

emotion that cannot disappear without an act of retribution.

4.3 Physiological Effects and Behavioral Responses

We observed above that *anger* and *rage* show some differences in the metaphors with which they are strongly associated. These differences appear to be compatible with the results of the significant metonymical collocates.

Table 4: Metonymical collocates of a*nger* and r*age*

Response	*anger* [18 collocates]	*rage* [13 collocates]
SCREAMING/ CRYING	*shout*_(S), *tear*_(S), *scream*_(V), *cry*_(S) [4 (22.2%)]	*scream*_(S), *cry*_(S), *scream*_(V), *weep*_(V), *tear*_(S), *cry*_(V) [6 (46.2%)]
REDNESS/ BODY HEAT	*livid*_(A), *flush*_(V), *darken*_(V), *dark*_(S) [4 (22.2%)]	*scarlet*_(A), *hot*_(A) [2 (15.4%)]
STIFFENING	*speechless*_(A), *taut*_(A), *stiffen*_(V), *tight*_(A) [4 (22.2%)]	*speechless*_(A) [1 (7.7%)]
AGITATION	*quiver*_(V), *tremble*_(V), *shake*_(V) [3 (16.7%)]	*tremble*_(V), *shake*_(V) [2 (15.4%)]
CONTORTION	*contort*_(V) [1 (5.6%)]	*contort*_(V) [1 (7.7%)]
Others	*pale*_(A), *bite*_(V) [2 (11.1%)]	*choke*_(V) [1 (7.7%)]

The most important difference is SCREAMING/CRYING FOR ANGER. Although SCREAMING is connected to FEAR, according to Kövecses (1990), it is also strongly connected to ANGER. Of *rage*'s significant metonymical collocates, almost half (46.2%) go into this category. Below are some examples (13a,b).

(13) a. I *scream* with rage. (HGN 3313)
b. Blindly, Alan ran to his room, where he beat and punched his bed and *cried* aloud in a rage like a child. (HJH 644)

The strong association between *rage* and SCREAMING/CRYING is compatible with the data on conceptual metaphors. As mentioned above, *rage* is strongly associated with the ANIMAL metaphor, which conceptualizes violent aspects of the emotion. As the ANIMAL metaphor is undoubtedly based on the angry act of screaming, it is

natural that *rage* is strongly associated with this behavioral response.

In contrast, *anger* is associated with different physiological and behavioral responses, such as AGITATION (14a), REDNESS (14b), and STIFFENING (14c).

(14) a. Suddenly Shiona was *trembling* with anger. (JXS 1586)
b. When he came back his face was *livid* with anger. (G04 2468)
c. She was almost *speechless* with anger now. (JY3 934)

With regard to STIFFENING, Kövecses (1990: 71) gives INABILITY TO MOVE ("*she was scared stiff*") and INABILITY TO SPEAK ("*I was speechless with fear*") as metonymies of FEAR. Kövecses does not state that these are metonymies of ANGER, but this study's results indicate that STIFFENING is also an important part of ANGER.

Finally, the data of this study show that *rage* is significantly associated with the verb *choke* as in (15).

(15) My voice came out like a croak — I was *choking* with rage. (BMS 1077)

This section examined the significant metonymical collocates. While *anger* and *rage* share quite many collocates, there are differences that should be noted. First, *rage* is more strongly associated with SCREAMING/CRYING than *anger*. This is compatible with the result that *rage* is more strongly associated with the ANIMAL metaphor. Secondly, *anger* is associated with different responses including AGITATION, REDNESS, and STIFFENING, while *rage* is significantly associated with INABILITY TO BREATHE. Metonymies such as STIFFENING/INABILITY TO BREATHE FOR ANGER have not been much discussed as metonymies of ANGER in previous studies.

5. Conclusion

This study investigated the similarities and differences in metaphors/metonymies between two near-synonyms of ANGER: *anger* and *rage*. With regard to emotion metaphors/metonymies, much discussion has concerned the following two issues: 1) Which conceptual metaphors/metonymies are mainly used for a particular emotion, and

2) which metaphors are the most central to the emotion. As for ANGER, a certain number of metaphors/metonymies have been proposed, and it is generally agreed that the central metaphor for ANGER is ANGER IS A HOT FLUID IN A CONTAINER. However, little attention has been paid to the similarities and differences between near-synonyms. As an exception, Turkkila (2014) compares near-synonyms of ANGER from the perspective of Issue 1 and concludes that near-synonyms of ANGER generally have the same metaphors.

Based on these discussions, the present study attempted to elucidate the similarities and differences of near-synonyms of ANGER from the perspective of the centrality of metaphors/metonymies to the emotion (Issue 2). To determine the centrality, this paper introduced the MI-score method and applied two criteria: 1) A collocate with a higher score is considered to be more central, and 2) the metaphors/metonymies that contain more significant collocates are more central. The analysis showed that ANGER IS A HOT FLUID IN A CONTAINER is the central metaphor for *anger*, whereas *rage* is strongly associated with ANGER IS A DANGEROUS ANIMAL, as well as with the FLUID metaphor. Above all, instantiations of the ANIMAL metaphor tended to top the list of *rage*'s significant collocates. Furthermore, through the categorization of the significant metaphorical collocates, it was found that *anger* represents different aspects of the emotion, while *rage* represents only the intense, violent aspects. This result is supported by the metonymical collocates. In this way, the method proposed in this article is a useful way of studying conceptual metaphors and metonymies.

Notes

1. It is thus important to distinguish conceptual metaphor from metaphorical expression. A metaphorical expression is a linguistic manifestation of a conceptual metaphor. In this article, a conceptual metaphor is called either conceptual metaphor or metaphor, and its linguistic instantiation is called metaphorical expression. Conceptual metaphors are written in capitals.
2. Turkkila's categorization of metaphors appears to be a little different from that of many previous studies, but we do not discuss their validity here. It is nevertheless obvious that the most frequent metaphors are generic-level metaphors.
3. The MI-score is the observed frequency divided by the expected frequency, converted to a base-2 logarithm (Hunston, 2002).

4. The collocates *sethe*_(V) and *seethe*_(V) should be counted as a single collocate *seethe*_(V).

An earlier version of this paper was presented at the 43rd Conference of Japan Association for English Corpus Studies, held at Kwansei Gakuin University in September 2017.

References

Akano, I. (2009). Corpus gengogaku [Corpus linguistics]. In K. Imai (Ed.), *Gengogaku no ryoiki 2* (pp. 125–148). Tokyo: Asakura Shoten.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29.

Deignan, A. (2005). *Metaphor and corpus linguistics*. Amsterdam: John Benjamins.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Kövecses, Z. (1990). *Emotion concepts*. New York: Springer-Verlag.

Kövecses, Z. (2000). *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge: Cambridge University Press.

Kövecses, Z. (2011). Methodological issues in conceptual metaphor theory. In S. Handl, & H. J. Schmid (Eds.), *Windows to the mind: Metaphor, metonymy and conceptual blending* (pp. 23–39). Berlin/New York: Mouton de Gruyter.

Krishnamurthy, R. (2003). *English collocation studies: The OSTI report*. Birmingham: University of Birmingham Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Oster, U. (2010). Using corpus methodology for semantic and pragmatic analysis: What can corpora tell us about the linguistic expression of emotions? *Cognitive Linguistics*, 21(4), 727–763.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stefanowitsch, A. (2004). Happiness in English and German: A metaphorical-pattern analysis. In M. Achard & S. Kemmer (Eds.), *Language, culture, and mind* (pp. 134–149). Stanford: CSLI.

Stefanowitsch, A. (2006). Words and their metaphors: A corpus-based approach. In A. Stefanowitsch & S. Th. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy* (pp. 63–105). Berlin: Mouton de Gruyter.

Suzuki, K. (2010). <Ikari> wo arawasu eigo ruigigo ga motsu metaphor eno senkou: Gainen hiyu riron no shiten kara [Preference of English emotional terms on metaphorical patterns]. *Kobe Papers in Linguistics*, 7, 60–77.

Turkkila, K. (2014). Do near-synonyms occur with the same metaphors: A comparison of anger terms in American English. *Metaphorik.de*, 25, 129–154.

（南澤　佑樹　大阪大学大学院　Email: un_permis_de_conduire@yahoo.co.jp）

「論文」

Move Development of London Hotel Overviews on Official Websites: Luxury Strategies in Overview Texts

Yukie KONDO

Abstract

This study examines the texts of 124 hotel overviews on the official websites of 3-5-star hotels in London, using move structure analysis, a method of genre analysis developed by Swales (1990). Despite the fact that readers of the hotel websites are neither "expert members" nor "members of the professional or academic community," which deviates from the existing established genre theory (Swales, 1990; Bhatia, 1993), discourse units, or moves and steps, were identified and hotel overviews had some aspects of a promotional genre in terms of communicative purposes. By developing those moves, typical structures of hotel overviews were found, and the implementation rate of the moves and keywords in each grade followed by a qualitative analysis of the sentences showed differences between lower-grade and higher-grade hotels in strategies to appeal to their potential guests. This study demonstrates how these differences are a result of the luxury strategy of "exclusivity" and "abstractness" adopted by higher-grade hotels.

1. Introduction

Since the Internet became available worldwide, hotels have had their own official websites, and online reservations have become prevalent. Online reservations have created an "unprecedented impact on travel and tourism, in general and in particular on hotel room bookings" (Law, 2009: 766). In the case of official websites, people expect to not only make a hotel room reservation but also to directly access information from the hotel. As websites are "an important means for a firm to communicate its service assurances with its consumers" (Chen and Dhillon, 2003: 311) and are "critical in the development of trusting relationships with customers" (Wang, Law, Guillet, Hung, and

Fong, 2015: 108), official websites can be the most reliable means of communication between hotels and the readers of the websites. Empirical studies on hotel websites have become specialized in areas such as design, usability, content, and quality (Law and Cheung, 2006; Choi, Letho, and Morrison, 2007; Chang, Kuo, Hsu, and Cheng, 2014), marketing activities (Li, Wang, and Yu, 2015), and users' purchase intention and attitude (Kim, Ma, and Kim, 2006; Li et al., 2015). However, there is still much to explore in terms of language use and discourse structures of websites. Official websites of hotels in the English language commonly have an "overview section" on the top page, wherein hotels briefly state what they have to offer or how they are attractive to potential guests. Since the appealing points are different among hotels, it can be presumed that the linguistic strategies adopted in the overview section are also different. Hotels are categorized into different grades, or stars, and there can be tendencies and/or differences in language use in each grade. The author is particularly interested in how higher-grade hotels express their "luxury" in their overview texts as a linguistic strategy. Considering the above points, this study investigates the linguistic strategies adopted by official websites of hotels in London to identify how they appeal to the readers of the websites. The major focus of this study is to investigate how hotel grades have an influence on language use; in particular, how higher-grade hotels' "luxury" appears in texts. This study analyzes their discourse structure through genre analysis using move structure analysis, proposed by Swales (1990). As hotel overviews are assumed to share a set of common communicative purposes, analysis of the texts using moves, which have "a local purpose" and also contribute to "the overall rhetorical purpose of the text" (Biber, Connor, and Upton, 2007), will contribute to revealing the typical structure of hotel overviews based on the shared communicative purposes they have.

Main efforts of this study are devoted to the development of the moves implemented in hotel overviews, as hotel overviews are treated as a genre for the first time as far as the author knows. The structure of this paper is as follows: Section 2 reviews the approaches of genre analysis; Section 3 explains the corpus compiled for this study as well as the methods used in this study; Section 4 describes how the author develops moves and steps; Section 5 conducts a quantitative analysis of the move implementation rate as well as keywords extracted in each grade. Section 6 discusses how linguistic strategies differ between higher- and lower-grade hotels by contrasting

the results acquired in Section 5 with sentence-by-sentence qualitative analysis, focusing on how higher-grade hotels use their luxury to appeal to the readers. Finally, Section 7 provides a summary and discusses limitation of this study and future research possibilities.

2. Genre analysis: Hotel overviews on websites as a genre

Although the term "genre" is widely used in various fields, such as art, music, and literature, genre analysis in this study is concerned with discourse classification focusing on language usage. Hyon categorized genre approaches into three: English for specific purposes (ESP), North American New Rhetoric studies, and Australian systemic functional linguistics (1996: 694). As this research treats hotel overviews as a "communicative event" between the hotels and the readers of their website, the ESP approach developed by Swales (1990) and later by Bhatia (1993) is adopted to apply move structure analysis.

Swales defines a genre as follows:

> A genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community, and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style. (1990: 58)

Bhatia, extending on Swales (1990), modifies the definition as follows:

> … it is a recognizable communicative event characterized by a set of communicative purpose(s) identified and mutually understood by the members of the professional or academic community in which it regularly occurs. Most often it is highly structured and conventionalized with constraints and allowable contributions in terms of their intent, positioning, form and functional value. These constraints, however, are often exploited by the expert members of the discourse community to achieve private intentions within the framework of socially recognized purpose(s). (1993: 13)

Swales (1990, 2004) and Bhatia (1993) explored academic and professional

genres such as research articles, sales promotion letters, job application letters, and legal texts within a "closed discourse community" constituted of expert members of that community, designating that as the characteristic of a genre. The discourse community of hotel overviews on websites is composed of the hotels that produce the websites, and the readers of the websites who seek information about the hotels; both parties share the communicative purpose of providing/receiving information about the hotel for future benefit. Although the readers of the websites are neither "expert members" nor "members of the professional or academic community," and in that sense, its discourse community is rather "open," hotel overviews have some aspects of promotional genres in that their communicative purposes include "capturing the attention of the potential customer," "offering and appraising the product or service in terms of the perceived interests, needs, or inhibitions of the potential customer," and also "must be short and effective" (Bhatia, 1993: 46). Websites run by business entities are essentially an arm of the advertising media that plays a role in turning passers-by into potential customers and, in turn, potential customers into actual customers. Toward that goal, hotel overviews are written in accordance with some conventions, if not constraints, "to achieve special effects or private intention" (Bhatia, 1993: 14).

3. Procedure

3.1 Corpus

The corpus for this study is composed of the texts of 124 overviews on the official websites of 3-5-star London hotels, collected between March and April, 2017. The texts that contained general remarks about the hotel, usually written on the top page or a relevant page such as a "home (page)," were defined as the overview. Table 1 shows the number of texts and tokens according to each grade.

Table 1: General information about the corpus

Hotel grade	Number of hotels	Tokens	Tokens Average	Tokens Minimum	Tokens Maximum
3-star	11	2414	219.45	31	659
4-star	66	10521	158.21	29	458
5-star	47	6289	132.55	21	348

The AA Hotel Guide 2016 (AA Publishing, 2015) was used to select the hotels of each grade in London. Since 2007, the AA and three other associations—VisitBritain, VisitScotland, and the Wales Tourist Board—agreed to follow the same criteria for grading based on the National Quality Assurance Standards (Baker, 2007). The new system divides hotels into five types: *Hotel*, *Town House Hotel*, *Country House Hotel*, *Small Hotel*, and *Metro Hotel*, according to location, the number of rooms, and the service. Most of the hotels came under the type *Hotel*, and therefore, *Hotel* was the only type treated in this study. Among the hotels, this study dealt with those having ratings of three stars or higher because only one hotel was ranked 2-star and zero hotels were 1-star.

3.2 Method

First, the texts of overviews on the official websites of 3-5-star hotels in London were collected to compile a corpus for analyses. In total, 11 overviews of 3-star hotels, 66 overviews of 4-star hotels, and 47 overviews of 5-star hotels were used. Hotel overviews usually contained the hotel name once or more than once in most cases. Considering that hotel names may affect the results of later analyses, all the hotel names were replaced by "HOTELNAME." Next, each overview was divided into several moves according to the function, or the communicative purpose, of each sentence. For this process, only one move was assigned to one sentence. When more than one function was found in a sentence, the function in the core of the sentence or the main clause of the sentence was examined. After moves were assigned to each sentence, all the sentences were tagged using CasualTagger[1]; in addition, typical language expressions in each move were extracted. To refine the moves, CasualConc[1] was used to search a move by a typical expression, or a typical expression by a move, and to confirm the integrity of moves and language expressions. Steps were also assigned when there were sub-divisions of a certain move. These processes of developing and refining moves were conducted by the author and two native English ESL teachers. As this study's main efforts were devoted to the development of the moves, rather than applying already developed moves to confirm the compatibility and reliability, we did not look into inter-rater reliability this time. Rather, we discussed to reach a consensus while refining moves. In the final procedure, keywords according to the hotel grades were derived using CasualConc. They were used as supplementary data to compare the linguistic characteristics of each grade.

4. Development of moves

4.1 Functions and communicative purposes

Before developing moves, the functions of the text were drawn out sentence by sentence. Table 2 shows examples of functions and sentences found in this process. Some sentences seemed to have several different functions; however, only one function was assigned to the core, or the main clause of the sentence.

Table 2: Examples of the functions and example sentences

Functions	Example sentences
Heading	About Us About the hotel
Definition	A beacon of British style and sophistication, HOTELNAME is a renowned five star hotel in central London.
Rhetorical description	HOTELNAME is a tranquil haven of peace and quiet amongst the bustle of central London.
Stating history	HOTELNAME was built in 1936 and was originally launched as The White House luxury apartments.
Stating architecture	It comprises thirteen Georgian Townhouses that date from 1731.
Stating location	HOTELNAME is located near Bond Street and Selfridges, offering London's best shopping areas right at your door step.
Stating facilities	The hotel features include functions rooms for up to 100 guests to suit your conference, meeting or wedding, as well as a hotel restaurant and bar.
Invitation	Come join us. Escape the city
Stating what the hotel does/provides	So whether you are planning a special event or to tie the knot in style, we promise it will be a very special occasion. From the moment you arrive, our goal is to make you feel at home.
Stating what the guests do/receive	On entering this charming hotel, you'll sense its unique atmosphere, making you feel at home. Whether travelling in or out of the capital, you are guaranteed a restful night's sleep
Stating next steps for potential guests	Book here for our Best Price Guarantee promise.
Welcome	Welcome to HOTELNAME We are delighted to welcome you to one of the most popular 5 star hotels in London…
Looking forward	We look forward to welcoming you.
Stating the hotel name and/or the manager's name at the end	MANAGERNAME, General Manager

Next, several functions were combined to identify communicative purposes as shown in Table 3 to assign moves later.

Table 3: Communicative purpose of each function unit

Communicative purposes	Functions
Heading	Heading
Defining the hotel	Definition Rhetorical description
Establishing features of the hotel	History Architecture Location Facilities
Establishing connections between the hotel and readers of the website	Invitation What the hotel does/provides What the guests do/receive Next steps for potential guests Welcome Looking forward Hotel name and/or the manager's name at the end

Heading

This unit always came at the beginning of the text or a paragraph, if at all. Not all the overviews had a heading, and on some websites, headings were separated from the overview texts because some headings were blended with other menus or icons. Other headings were just the name of the hotel, and/or the address and telephone number, while other headings had particular functions.

Defining the hotel

This unit usually came after the *Heading* or at the beginning if there was no *Heading*. In this unit, a hotel presented the readers of the website with its definition by declaring what star rating it had and/or by illustrating the hotel with appropriate rhetorical expressions or metaphors.

Establishing features of the hotel

This unit usually came after the *Defining the hotel*. In this unit, the hotel introduced its specific features, such as history, architecture, location, and facilities. As for the location, this move gave specific information about where the hotel was located by stating the address, place, or street name; or landmarks near the hotel or how long it would take to get from the hotel to the landmarks. The history provided information about how long it had been in operation, as well as when the hotel opened, and/or its background. It was also sometimes combined with information about the architecture. The number and/or sort of rooms, restaurants, bars, and/or services such as Internet connection or air-conditioning, were stated in the facilities unit. The information about

the staff was also described in this unit.

Establishing connections between the hotel and readers of the website

A closer examination of the function units that were not classified into any of those mentioned above revealed that they served to establish connections between the hotel and the readers of the website. When this unit came at the beginning of the overview, it established this connection by inviting the reader, greeting them, gaining attention, and so forth. When it came at the end of the overview, it gave a final comment from the hotel to the readers by describing features of the hotel and/or stating for whom and for what the hotel is suitable.

4.2 Developing and refining moves

The moves were developed according to the communicative purposes mentioned in the earlier section. In the case when a move was related to several functions, the move was broken down and steps were assigned under the move. Consequently, four moves were developed and named *Move 0: Heading, Move 1: Defining self*, *Move 2: Establishing features*, and *Move 3: Establishing connections*; three steps in *Move 2* were also developed and named *Step 1: History/architecture*, *Step 2: Location*, and *Step 3: Facilities*. There were only a few descriptions of architecture, and they were always accompanied by descriptions of the history; hence, history and architecture were treated in the same step. Move and step numbers were allotted according to the most common order in which they appeared in each overview. The moves and steps developed are shown in Table 4.

Table 4: Moves and steps

Moves	
Move 0:	Headings
Move 1:	Defining self
Move 2:	Establishing features
	Step 1: History/architecture
	Step 2: Location
	Step 3: Facilities
Move 3:	Establishing connections

Headings holding communicative purposes of another move were categorized in that move in accordance with the communicative purpose. [1], [2], and [3] shown

below are examples of headings. [1] is a heading, but simultaneously it serves to define the hotel; therefore [1] was considered as *Move 1: Defining self*. [2] describes the location of the hotel; therefore, it was considered as *Step 2: Location* (of *Move 2: Establishing features*). [3] does not have any other communicative purposes other than to serve as a heading, and hence, it was purely considered as *Move 0: Heading*. *Move 0: Heading* was not analyzed this time.

[1] A STYLISH HOTEL IN LONDON
[2] Hotel in Kensington
[3] HOTEL OVERVIEW

Table 5 shows an example of an overview that contains all moves, and Table 6 shows an example that contains only one move. As shown in Table 5, some moves occurred repeatedly in some cases.

Table 5: Example of a hotel overview containing all moves

Move	Text
Move 1: Defining self	ONE OF THE FINEST 5 STAR HOTELS CHELSEA HAS TO OFFER HOTELNAME is the epitome of classic elegance; a fine example of a five star hotel, Chelsea – London's finest. Built in 1890, HOTELNAME is one of the finest and most elegant five star hotels in London.
Move 2: Step 2: Location	This beautiful, lovingly restored red-brick Edwardian hotel is situated just around the corner from Chelsea's fashionable Sloane Square and the beautiful borough of Kensington, making it the perfect location to explore everything London has to offer.
Move 1: Defining self	HOTELNAME offers luxury accommodation with traditional old age elegance and quintessentially British luxury within a peaceful city retreat.
Move 2: Step 3: Facilities	Personal touches such as complimentary tea and homemade biscuits at 4:00pm, champagne at 6:00pm, hot chocolate and biscuits at 9:30pm and an honesty bar in the hotel lounge add to HOTELNAME's wonderfully unique charm.
Move 3: Establishing connections	With staff on hand to satisfy your every whim, HOTELNAME will become your very own indulgent and luxurious home away from home.
Move 2: Step 3: Facilities	As a guest at the luxury HOTELNAME, you'll enjoy rest and respite in a choice of 35 luxurious, tasteful and individually decorated boutique suites and rooms. Steeped in Edwardian splendour, each room is theatrically themed and adorned with Victorian antiques. Most of the suite rooms have their own cosy working fireplace and many offer a view of the tranquil private Cadogan garden to which guests have exclusive access.
Move 3: Establishing connections	Book your stay at one of the most luxurious 5 star hotels in Chelsea, London.

Table 6: Example of a hotel overview containing only one move

Move	Text
Move 1: Defining self	Surrounded by greenery in the heart of Mayfair, the luxurious and redesigned HOTELNAME is unique in London, yet still true to the city.

As mentioned earlier, there were some cases where one sentence had several functions; in those cases, only one move was assigned to the core, or the main clause of the sentence. For example, the sentence in Table 6 describes the location in the beginning of the sentence, but the core of the sentence defined the hotel using rhetorical expression; therefore, it was considered as *Move 1: Defining self*.

5. Results

5.1 Move frequencies

Move frequencies and implementation rates are shown in Table 7 and Figure 1. The move with the highest implementation rate was *Move 2: Establishing features* (91.1%), followed by *Move 1: Defining self* (83.1%). It was common for hotel overviews to define their hotels and establish their features such as location, facilities, or history/architecture. *Move 3: Establishing connections* (66.9%) was found not to be as common a move.

Table 7: Move frequencies

Number of overviews	124
Number of overviews that have Move 1: Defining self	103 (83.1%)
Number of overviews that have Move 2: Establishing features	113 (91.1%)
Number of overviews that have Move 3: Establishing connections	84 (67.7%)

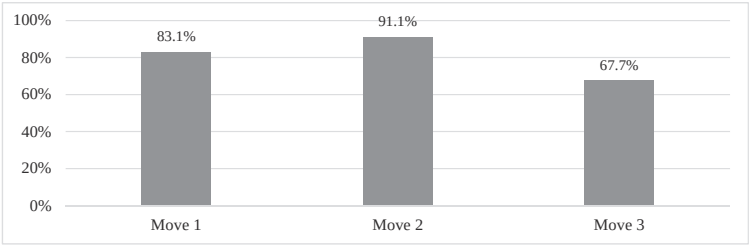


Figure 1: Implementation rate of each move

Move frequencies and implementation rates according to the grade of the hotels are shown in Table 8 and Figure 2. The move with the highest implementation rate was *Move 2: Establishing features*: 100% by 3-star hotels, 92.4% by 4-star hotels, and 87.2 % by 5-star hotels. *Move 1: Defining self* was implemented a little more than 80% by all the grades. Establishing their own features was crucial for lower-grade hotels. A clear difference among grades was found in the implementation rate of *Move 3: Establishing connections*. It was implemented by more than 81.8% of 3-star hotels and 75.8% of 4-star hotels, but only by 53.2% of 5-star hotels.

Table 8: Move frequencies according to hotel grades

	3-star	4-star	5-star
Number of overviews	11	66	47
Number of overviews that have Move 1: Defining self	9 (81.8%)	55 (83.3%)	39 (83.0%)
Number of overviews that have Move 2: Establishing features	11 (100%)	61 (92.4%)	41 (87.2%)
Number of overviews that have Move 3: Establishing connections	9 (81.8%)	50 (75.8%)	25 (53.2%)

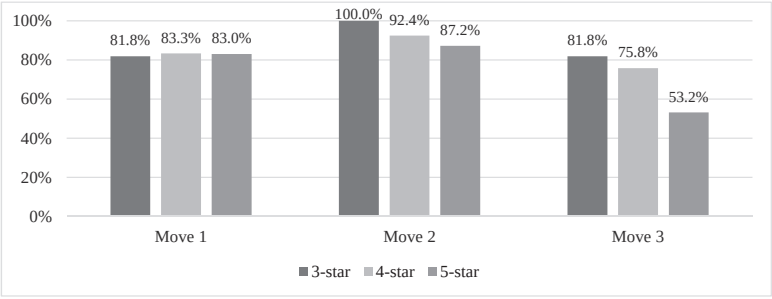


Figure 2: Implementation rate of each move according to hotel grades

Move 2: Establishing features was the only move that had steps, and the implementation rates of steps were different among each grade as shown in Table 9 and Figure 3. *Step 2: Location* and *Step 3: Facilities* were implemented by more than 70% of 3- and 4-star hotels, but the implementation rates by 5-star hotels were lower in both steps. *Step 1: History/architecture* showed a different tendency, implemented by much fewer hotels, 27.3% by 3-star, 10.6% by 4-star and 29.8% by 5-star hotels. *Step 1* was the only step that had the highest implementation rate by 5-star hotels.

Table 9: Implementation rate of each step in Move 2 according to hotel grades

	3-star	4-star	5-star
Number of overviews	11	66	47
Number of overviews that have Step 1: History/architecture	3 (27.3%)	7 (10.6%)	14 (29.8%)
Number of overviews that have Step 2: Location	8 (72.7%)	47 (71.2%)	25 (53.2%)
Number of overviews that have Step 3: Facility	8 (72.7%)	54 (81.8%)	31 (66.0%)

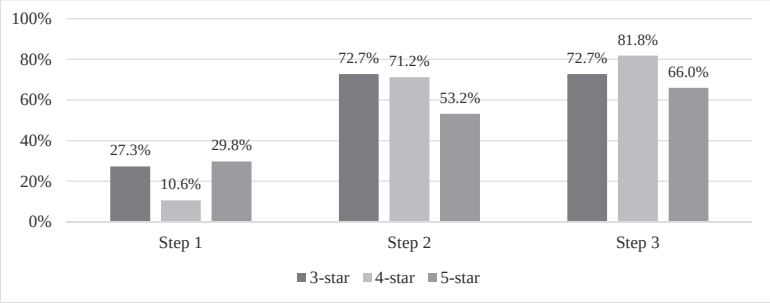


Figure 3: Implementation rate of each step in Move 2 according to hotel grades

5.2 Keywords

To examine the characteristics of words used in each grade, keywords were derived according to the value of log-likelihood using CasualConc and shown in Table 10. The hotel overview corpus as a whole was used as a reference corpus. It was evident that in 5-star hotels, many words referring to qualities such as "luxury," "5-star," "luxurious," "English," "private," and "British," were ranked. 4-star hotels also had those words referring to qualities: "4-star," and "convenient," as did 3-star overviews: "value" and "superb"; however, the number was smaller. Words referring to a place or location were observed through all the grades: "Wembley," "stadium," "arena," and "Ruislip" in 3-star hotels and "Kensington," "station," "airport," "link," "transport," and "Victoria" in 4-star hotels; however, there was only one, "Mayfair," that ranked in 5-star hotels. The words related to location seen in 3- and 4-star hotels were also regarded to be facilities outside the hotel. On the other hand, 5-star hotels had words referring to the hotel's own facilities: "suite" and "spa." In summary, there were more abstract qualities and hotels' own facilities stated in higher-grade hotels'

overviews, while more concrete information about facilities outside the hotels appeared in lower-grade hotels.

Table 10: Keywords according to hotel grades

	3-star Keywords	Log-likelihood	4-star Keywords	Log-likelihood	5-star Keywords	Log-likelihood
1	Wembley	111.60	4-star	3.94	luxury	20.35
2	stadium	35.90	shop	2.24	5-star	13.21
3	value	18.97	from	2.14	Mayfair	9.03
4	arena	18.80	leisure	1.96	suite	8.68
5	tour	9.87	Kensington	1.90	spa	7.91
6	event	8.02	station	1.87	in	7.82
7	you	7.40	airport	1.68	one	7.16
8	Ruislip	7.30	convenient	1.68	star	5.92
9	choice	7.15	link	1.61	luxurious	5.07
10	200	6.27	transport	1.61	English	5.05
11	centrally	6.27	Victoria	1.44	private	5.05
12	superb	6.11	point	1.43	British	4.70

6. Discussion

In this section, the author discusses how hotel grades influence overview texts, particularly how higher-grade hotels' "luxury" appears in the form of language. Contrasting the results of the move implementation rate with a qualitative analysis of each sentence, it was found that "luxury" was expressed in hotel overviews by incorporating "exclusivity" and "abstractness" in the texts.

6.1 Luxury strategy in overview texts: Exclusivity

The major difference in the move implementation rates was found in that of *Move 3: Establishing connections*. In this move, "you" was a high ranked keyword as shown in Table 10. From qualitative analyses of the sentences in this move, it was found that second person pronouns as well as imperative forms were used to call for action. [4] is encouraging potential guests to stay at their hotel using the imperative form, and [5] explains how valuable the hotel stay will be for the guests using "perfect

choice." [6] refers to various objectives for guests and "special deals" to appeal to potential guests with different needs.

[4] Stay at our HOTELNAME for a royal trip you will never forget. (3-star)

[5] Making it the perfect choice if you are looking for Wembley hotels. (3-star)

[6] Whether you are looking for a perfect weekend break, romantic two-night stay or seasonal getaway, you can check out our special deals and find something that suits you. (4-star)

The reason why 5-star hotels used this move to establish a connection with the readers of the website much less than lower-grade hotels did can be explained by the luxury strategy of making "exclusivity." Phau and Prendergast argue that luxury brands compete based on the ability (1) to evoke exclusivity, (2) to have a well-known brand identity, (3) to increase brand awareness and perceived quality, and (4) to retain sales levels and customers' loyalty (2000: 124). By reaching potential guests through official websites, hotel brands can achieve the second and the third points, but if they establish connections by using *Move 3: Establishing connections*, they might fail to evoke exclusivity as the readers might feel that the hotel is available for anyone. As Kapferer and Bastien indicate, a "luxury product can communicate via the internet, but should not be sold there" (2009: 207). 5-star luxury hotels make themselves attractive as a luxury hotel and increase awareness and perceived quality, but do not necessarily try to sell themselves to everyone who reads their websites. By not directly addressing the readers of the website to establish connections, 5-star hotels can maintain their exclusivity.

6.2 Luxury strategy in overview texts: Abstractness

Differences were also observed in the move implementation rates of *Move 2: Establishing features*. The implementation rate gradually declined as hotel grades rose: 100% for 3-star hotels, 92.4% for 4-star hotels, and 87.2% for 5-star hotels. However, even with a lower implementation rate of *Move 2*, higher-grade hotels did mention features such as location and facilities. The difference was that they did so not in the

main clause but in modifying clauses or phrases of *Move 1* sentences.

Example sentences [7] and [8] are those of *Move 2* and [9] and [10] are those of *Move 1*.

Example sentences in *Move 2: Establishing features*

[7] Located in Ruislip HOTELNAME is within 1 mile of the A40 and M25 motorway (junct 16) and within a short drive of the business towns of Uxbridge, Harrow, Greenford and Watford. Heathrow Airport is reached within a 20 minute drive. (3-star)

[8] Our Umami Restaurant features food inspired by oriental cuisine comprised of noodles, tapas, soup and grilled dishes. (4-star)

In [7], information on location and facilities are mentioned in the core of the sentences as well as in the participle phrases. [8] does not have participle phrases and the entire sentence explains the restaurant and its food.

Example sentences in *Move 1: Defining self*

[9] Superbly located on a quiet Mayfair square, HOTELNAME recaptures the spirit of old-world luxury. (5-star)

[10] With world-famous restaurants and a stunning spa, we offer a fashionable and timeless base in the centre of the British capital. (5-star)

In [9] and [10], the core of the sentences falls under *Move 1: Defining self*, but the sentences also have information on the location or facilities in their modifying phrases. (Modifying phrases are underlined by the author.) The participle phrase of [9] explains the location, while the sentence defines the hotel using the abstract expression "the spirit of old-world luxury." The prepositional phrase of [10] talks about the hotel's facilities such as a restaurant and a spa, while the sentence also defines the hotel with an abstract concept, "a fashionable and timeless base." [10] also has information on the location, but it only mentions "the centre of the British capital"; hence, it still gives an abstract sense to the readers of the website.

Hansen and Wänke find that abstract product descriptions are considered as more luxurious than concrete product descriptions, and advertisers tend to use more abstract language when they describe their luxury products (2011: 794). Overviews of higher-grade hotels also use abstract descriptions in this way. Using abstract expression in the core of the sentence and concrete information in modifying phrases, they can make their description sound luxurious while giving specific features of the hotels.

Finally, the reason why *Step 1: History/architecture* was preferred by 5-star hotels can also be explained by the strategy of using "abstractness" in overviews. A hotel's history and architecture can be considered as a quality related to the "dream value" of luxury (Dubois and Paternault, 1995: 70). Chandon, Laurent, and Valette-Florence discuss how luxury brands can use the Internet, while maintaining the "dream value of luxury," by providing consumers with "such stuff as dreams are made of" (2016: 301). They take "brand history and heritage, creation legends, or information about exceptional quality craftsmanship and materials" (2016: 301) as examples of what dreams are made of. As 5-star hotels are regarded as luxury hotels, stating abstract value such as history and architecture in their overviews rather than concrete information can be regarded as expressing their extravagance in language, which can be more appealing to those readers who are looking for luxury.

7. Conclusion

Hotel overviews appeared to be freely created texts; however, by developing moves and analyzing the texts on the basis of move usage, a typical construction of hotel overviews was found. Tendencies and differences were observed among different hotel grades. Defining the hotel and establishing its features were two crucial pieces of information in overviews for both higher- and lower-grade hotels; however, establishing connections between the hotel and readers of the websites was less crucial, especially in higher-grade hotels. The reason for these differences can be explained by the luxury strategy that higher-grade hotels take. By not directly addressing or inviting the readers, higher-grade hotels can maintain exclusivity, thus expressing their luxury even in the overview texts. Another luxury strategy that higher-grade hotels adopt is abstractness; they give specific information in modifying phrases and keep the core of the sentence abstract. Using luxury strategy is also found in the higher implementation

rate of *Move 2 Step 1: History/architecture* by 5-star hotels. Referring to the history and/or architecture in the overview, hotels can include the "dream value" of luxury in the texts. To sum up, differences in appeal to the readers among different grades are due to the fact that higher-grade hotels apply luxury strategies in their overview texts.

The findings mentioned above were obtained by compiling a corpus of hotel overviews and treating them as a genre. Hotel overviews as a genre can be an example of a genre that has an open discourse community but still has shared communicative purposes and therefore a typical move structure. Now that the moves for hotel overviews have been developed, further research can be conducted to investigate lexico-grammatical features and text patterns in different moves and/or grades. In this study, only the keywords in each grade were discussed to see the differences among hotel grades. In future studies, it would be interesting to extract keywords from each move, or to carry out another corpus-based analysis. A corpus-based analysis such as correspondence analysis could reveal other characteristics between the different grades of hotels, or it could discover other factors that differentiate hotels based on linguistic features instead of the existing hotel star grading system. Furthermore, another corpus of hotel overviews in other locations should be compiled to see whether the move structure found in this study is exclusive to hotels in London or if it can be applied to hotels in general, as the corpus compiled in this study was rather small and not well-balanced due to the limited number of hotel overviews. Despite these limitations, this corpus-based study sheds new light on how "luxury" appears in texts and how the luxury strategy is used in the form of language. This suggests that corpus-based move structure analysis enables researchers to find strategies used in a seemingly unconstrained discourse that has yet to be regarded as a genre.

Acknowledgments

The author wishes to thank Dr. Atsuko Umesaki and anonymous reviewers for their constructive comments that improved an earlier version of this manuscript.

Note

1. CasualTagger and CasualConc are freeware concordance programs for Macintosh OS X created by Dr. Yasuhiro Imao of Osaka University. The programs can be downloaded from https://sites.google.com/site/casualconc/

References

AA Publishing. (2015) *The AA Hotel Guide 2016*. Basingstoke: Author.

Baker, B. (2007, January 11) "New UK Hotel Rating System Goes Public." *The Guardian*. Retrieved from https://www.theguardian.com/travel/2007/jan/11/travelnews.uk.hotels

Bhatia, V. (1993) *Analysing Genre: Language Use in Professional Settings*. London: Longman.

Biber, D., U. Connor and T. A. Upton (2007) *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*, Amsterdam: John Benjamins Publishing.

Chandon, J. L., G. Laurent and P. Valette-Florence (2016) "Pursuing the Concept of Luxury: Introduction to the JBR Special Issue on 'Luxury Marketing from Tradition to Innovation.' " *Journal of Business Research 69*, 1: 299–303.

Chang, K. C., N. T. Kuo, C. L. Hsu and Y. S. Cheng (2014) "The Impact of Website Quality and Perceived Trust on Customer Purchase Intention in the Hotel Sector: Website Brand and Perceived Value as Moderators." *International Journal of Innovation, Management and Technology 5*, 4: 255–260.

Chen, S. C. and G. S. Dhillon (2003) "Interpreting Dimensions of Consumer Trust in E-commerce." *Information Technology and Management 4*, 2–3: 303–318.

Choi, S., X. Y. Letho and A. M. Morrison (2007) "Destination Image Representation on the Web: Content Analysis of Macau Travel Related Websites." *Tourism Management 28*, 1: 118–129.

Dubois, B. and C. Paternault (1995) "Understanding the World of International Luxury Brands." *Journal of Advertising Research 35*, 4: 69–76.

Hansen, J. and M. Wänke (2011) "The Abstractness of Luxury." *Journal of Economic Psychology 32*, 5: 789–796.

Hyon, S. (1996). "Genre in Three Traditions: Implications for ESL." *TESOL Quarterly 30*, 4: 693–722.

Kapferer, J. N. and V. Bastien (2009) *The Luxury Strategy: Break the Rules of Marketing to Build Luxury Brands*. London: Kogan Page.

Kim, W. G., X. Ma and D. J. Kim (2006) "Determinants of Chinese Hotel Customers' E-satisfaction and Purchase Intentions." *Tourism Management 27*, 5: 890–900.

Law, R. (2009) "Disintermediation of Hotel Reservations." *International Journal of Contemporary Hospitality Management 21*, 6: 766–772.

Law, R. and C. Cheung (2006). "A Study of the Perceived Importance of the Overall Website Quality of Different Classes of Hotels." *International Journal of Hospitality Management 25*, 3: 525–531.

Li, X., Y. Wang and Y. Yu (2015) "Present and Future Hotel Website Marketing Activities: Change Propensity Analysis." *International Journal of Hospitality Management 47*: 131–139.

Phau, I. and G. Prendergast (2000) "Consuming Luxury Brands: The Relevance of the 'Rarity

Principle.' " *Journal of Brand Management 8*, 2: 122–138.
Swales, J. (1990) *Genre Analysis: English in Academic and Research Settings*. New York: Cambridge University Press.
Swales, J. (2004) *Research Genres: Explorations and Applications*. Cambridge: Cambridge University Press.
Wang, L., R. Law, B. D. Guillet, K. Hung and D. K. C. Fong (2015) "Impact of Hotel Website Quality on Online Booking Intentions: ETrust as a Mediator." *International Journal of Hospitality Management 47*: 108–115.

（近藤　雪絵　立命館大学　Email: kondoyu@fc.ritsumei.ac.jp）

「論文」

日本人中高校生の英作文における複合動詞
― ゼロ動詞派生名詞とその対となる動詞の観点から ―

山本史歩子

Abstract

The aim of this paper is to examine the frequency of zero-deverbal nouns (e.g., a look, a visit, etc.), their corresponding verbs (e.g., to look, to visit, etc.), and composite predicates (e.g., take a look, make a visit, etc.) in the English composition of junior and high school students based on the Japanese EFL learner corpus (JEFLL Corpus). Further, the use of them in the students and that in native speakers of English taken from Wordbanks online are compared.

This paper has revealed that Japanese students had considerable difficulties in understanding the parallel between zero-deverbal nouns and their corresponding verbs, which could be one of the crucial factors in the extremely low frequency of use in composite predicates in their writings. In addition, they showed a lack of knowledge of verbs in composite predicates which suffer semantic bleaching and thus come to function as filling a syntactic slot.

In conclusion, it is desirable to make the student aware of the structure and productivity of composite predicates through analogy and inferencing.

1．はじめに

1.1　目的

日常的に大学生の英作文を添削する教員の多くは，読解力の高い学生でもいざ書くということになると，読むこととの間にかなりの隔たりがあると感じているのではないだろうか。具体例を挙げるならば，学生たちは関係副詞や分詞構文等を解釈するのにさして苦労はしないが，それらの構文を英作文では避ける傾向が観察される。英作文は理解構文と使用構文の差が如実に反映される。

構文以外にも注目すべき点はコロケーション，特に複合動詞（e.g., take a look, make a visit, etc.）の使用頻度の低さである。山本（2017）では，日本人大学生と英語を母語とする大学生のエッセイにおける複合動詞・句動詞・複合前置詞（e.g., in terms of, in (with) relation to, etc.）の使用頻度を分析した結果，句動詞に関しては日本人大学生の方が高い使用頻度が確認されたが，複合動詞と複合前置詞に関しては，英語母語話者の約半分であったことが報告されている。

本稿の目的は，JEFLL Corpus に基づいて日本人中高校生の英語学習者を対象に複合動詞を使用するための前提となる動詞派生名詞（e.g., a look, etc.）とその対となる動詞（e.g., to look, etc.）及び複合動詞（e.g., take a look, etc.）の認識の程度，使用パタン及び頻度を検証する。更に，Wordbanks online を用いて英語母語話者との使用パタンと頻度を比較し，最後に複合動詞の使用頻度の低さの要因について考察をする。本稿で扱うコーパスについては，3.1 節のデータと分析方法で再度言及をする。

1.2 複合動詞

複合動詞は，形式上「基本動詞 (do, give, have, make, take) +（不定冠詞）+ 動詞派生名詞 +（前置詞）」から構成されている（e.g., take care of, make a wish, give a cry, etc.）（Nickel,1968; Dixon, 1991; 相沢, 1999; Brinton and Akimoto, 1999）。

基本動詞は統語上動詞の位置を占める文法的機能を果たし，その意味は動詞派生名詞が担う。動詞派生名詞については，動詞と同形のゼロ派生名詞（e.g., make a guess/guess），音声上関係している名詞（e.g., give advice/advise），派生接辞が付く名詞（e.g., have an argument/argue），語源的に関係する名詞（e.g., have a thought/think）まで幅がある（Brinton and Akimoto, 1999: 2）。基本動詞と動詞派生名詞は必須項目であるが，不定冠詞と前置詞の有無は任意である。しかしながら，複合動詞は不定冠詞が動詞派生名詞に先行する形式が一般的とされている。複合動詞の構造はいたってシンプルであり，その意味は動詞派生名詞から推論可能である点において，英語学習者の使用を妨げる要因が一見すると認められない。

2．先行研究

複合動詞は通時的（秋元（編）, 1994; Brinton and Akimoto, 1999; Claridge, 2000），共時的（Nickel, 1968; Live, 1973; Nunburg et.al., 1994; 相沢 , 1999）の両

側面から多くの貴重な研究がなされている。Brinton and Akimoto (1999: 17–18) は，通時的観点から複合動詞の発達は名詞の抽象化に伴う脱範疇化，統語的再分析（動詞と名詞が 1 つのユニットとして機能）などのプロセスを経ていることから文法化と捉えているが，Traugott (1999: 257–259) は，文法化は内容語（名詞，動詞，形容詞など）が文法的機能を果たす方向へと発達することから，複合動詞の発達は語彙化であり，その後にイディオム化した例としている。共時的観点から，相沢（1999: 210–211）は，複合動詞の使用域に着目をし，新聞・小説・学術書における複合動詞の動詞を調べた結果，学術書では知的活動や思考による伝達行為を表す名詞と結びつく make 複合動詞の使用頻度が高かった。一方，主観的で日常生活の活動を表す名詞と結びつく have 複合動詞は極めて低い頻度であった。同じ複合動詞でも使用域が異なることを提示している。

単純動詞には見られない複合動詞の特徴は複数認められる。1 つは，単純動詞の不在（e.g., do homework, make an effort, etc.）（Quirk et al., 1985: 751–752）。次に，複合動詞の相的機能（aspectual function）。複合動詞を構成する不定冠詞 a/an の介在によりゴールが暗示された 1 度きりの出来事・行為，換言すれば，atelic な行為（終点は含意されない）を telic な行為（完遂・完了）へと変えることが可能になることを示唆している（e.g., move/make a move）（Live, 1973; Wierzbicka, 1984; Brinton and Akimoto, 1999）。次に，英語に特徴的な文末重点の観点から，Quirk et al.（1985: 1401–1402）は，動詞によっては SV（e.g., ‘she replied.’) だけで構成される文は稀であり，それを補う手段の 1 つとして複合動詞（e.g., ‘she made a reply.’）を用いるのがより一般的としている。最後に，Koskenniemi（1977: 83–84）は，弱強詩脚の韻律（iambic meters）を用いるドラマや無韻詩（blank verse）では，文末に複合動詞を置くことで（強勢を受ける名詞を置くこと），文に快音調的なリズム（euphonic cadences）を生むと論じ単純動詞にはない複合動詞の優位性を主張している。

他にも，複合動詞の統語上の柔軟性が認められる。Brinton and Akimoto（1999: 2–3）は，修飾語句の挿入の容易性（e.g., make a big impact vs. *impact bigly）を挙げている。Nickel（1968: 15–16）は，関係代名詞の先行詞への取出し（e.g., ‘*The grin she gave* with this remark nearly put Dixon right off his stroke.’）を挙げ，複合動詞の “descriptive force”（記述力）が使用増加の動機の 1 つになっていると述べている。また，Nunburg et al.（1994: 520）も，複合動詞の統語的柔軟性を示す二重受動化（e.g., ‘Advantage was taken of the students.’, ‘The students were taken

advantage of.’）に言及している。

英語教育におけるコロケーション習得の重要性についても多くの重要な研究がなされている（Sinclair, 1991; Lewis, 2000; De Cock, 2000; 堀，2009; 小屋，2015）。Sinclair（1991: 109–110）は，英語のテキストを読み込む際に 2 つの原則（the open-choice principle と the idiom principle）が働くが，基本的にはイディオムの原則が最初に働き，それが不適切と認識されると自由選択原則に基づいて適切に解釈されるとしている。ゆえに，英語母語話者の記憶には，膨大な量の ‘semi-preconstructed phrase’ 言わば「半固定化したフレーズ」が貯蔵されており，必要に応じて即座にそのフレーズを組み合わせて理解するので，独立した語よりもイディオム，チャンクの解釈が優勢になり，多くの語はコロケーションの形で使用されることを指摘している。

英語学習者にコロケーションを教えることの利点として，Hill（2000: 53）は，英語母語話者の言語活動の 70% 以上が何等かの形のコロケーションで処理されていること，堀（2009: 26–27）は，語は単独では用いられず他の語との関係で意味が決定されていること，Nation（2001），Nesselhauf（2003）は，英語母語話者のような流暢な言語活動の実現には欠かせないことなどの観点から，英語学習者にコロケーションを教えることが英語学習の有効的な手段の 1 つであると論じている。更に，Hill（2000: 54–55），Nation（2003: 320）は，英語母語話者は発話の際や聞くとき一語一語で認識をしているのではなく，チャンクで認識をして反応することにより，速いスピードでの言語処理が可能になることを指摘している。

また，教室で複合動詞を教える際のコロケーションリストの作成は日常的に使用されるものを提示することが重要とされているが（Lewis, 2000: 167; Nation, 2003: 328, 335），小屋（2015: 40–42）によれば，複合動詞を含めたコロケーションを詳細に解説している英語の教科書はほとんどなく，教科書によって扱っているコロケーションも異なる。加えて，教科書に掲載されている複合動詞と英語母語話者が日常的に使用する複合動詞とは使用パタンにかなりの隔たりがあることが報告されている（Koya, 2004）。

複合動詞に関する研究は様々な領域でなされているが，動詞と動詞派生名詞の使用頻度を出発点とした複合動詞の研究はほとんどなされていない。

3．分析

3.1　データと分析方法

本稿では，JEFLL Corpus をもとに頻度分析を行う。JEFLL Corpus は，日本人の中学生と高校生約 1 万人が書いた英作文を収集したコーパスで，規模としては 669,281 語を収録している。サブコーパスとして，中学生と高校生の別，学年・国立・公立・私立の別，各学校の偏差値の別，テーマ別など詳細に絞ることができるが，全例を分析対象とする。[1] JEFLL Corpus の比較対象となる英語母語話者のデータは, Wordbanks online から抽出をする。Wordbanks online は，Harper Collins 社が作成した Bank of English（約 6 億 5 千万語を収録した最大規模のコーパス）の内公開されている 102,813,738 語からなるコーパスである。

JEFLL Corpus は日本人中高校生の英語学習者コーパスであるので，比較対象となる英語母語話者のコーパスは，英語を母語とする大学生のエッセイを収録した LOCNESS（the Louvain Corpus of Native English Essays）が考えられるが，学習者コーパスは, 母語話者でも非母語話者であってもテーマや書いた環境（辞書の有無，制限時間など）によって使用される語彙や構文が左右される問題を抱えている（Biber and Reppen, 1998; 飯尾，2013）。Aijmer（2002: 63–65）は，イギリス英語を収録している LOB Corpus（the Lancaster-Oslo/Bergen corpus）と LOCNESS を基に法助動詞 must の根源的使用（deontic use）と認識様態的使用（epistemic use）の使用頻度を分析した結果，前者では根源的使用が認識様態的使用の 2 倍確認されたが，後者では両使用は同程度であったことを報告している。つまり，英語母語話者であっても，学習者コーパスの場合必ずしも一般的な傾向を反映しているとは一概には言えない。[2] また，『中学校学習指導要領解説　外国語編』（2017: 93）では，「実際の言語の使用場面や言語の働き」などを考慮した教材の創意工夫が求められている。従って，本稿では複合動詞の一般的な使用状況を提示するために，Wordbanks online の usbooks（fiction & non-fiction, 1988–1995, 5,410,682 語）を使用する。Wordbanks online のサブコーパスには他に新聞・パンフレット類・雑誌があるが，これらのジャンルでは本が最も中道的であると言える。[3]

本稿では，データが日本人中高校生による英作文であること，動詞と同形の動詞派生名詞は単音節から成る単純な語であることが多いこと，同形の使用比率の検証の重要性を考慮し，ゼロ動詞派生名詞（以下動詞派生名詞とする）を構成素とする複合動詞を分析の対象とする。動詞派生名詞は相沢（1999: 86–87）

が代表的とする複合動詞の中から，中学生で習得する語を選択した（scream と escape は高校生で習得する語）。次例参照。

(1) a. 身体活動：give a cry, have a laugh, give a scream, give a shout, have a smile
b. 身体動作：have/make an escape, take/have a look, make a stop, make a visit, take/have a walk
c. 知的活動：give an answer, have/make/take a guess, give an order, give/make a report, have a talk
d. 意志行為：take care (of), have a hope, make a promise, give/have a try, make a wish

一般的に，類義関係にある動詞派生名詞は同じ動詞と共起する傾向が認められており，プロトタイプ的な複合動詞の習得が周辺的な複合動詞の習得を容易にすると言われている。例えば, take medicine から take a pill，take a tablet（Nation, 2001: 325–328）。

3.2 身体活動 : cry, laugh, scream, shout, smile

最も基本的な身体活動を表す動詞派生名詞，単純動詞及び複合動詞の頻度分析の結果を表 1 に示す。D-Noun（Deverbal Nouns）は動詞派生名詞を，CP（Composite Predicates）は複合動詞を指す。身体活動を表す動詞派生名詞は，have や give と共起することが多く観察される。

注目すべき点は，英語母語話者と比較して日本人中高校生は動詞派生名詞の使用が極端に低く，scream は 5%，laugh は 3%，cry と shout にいたっては 1 例も確認されていない。ただし，例外的に smile だけが 39%（D-Noun と CP との合算）と高頻度で使用されているが，動詞派生名詞の smile という認識よりむしろ日本語に「カタカナ英語」として定着している「スマイル」からの想起であると推察される。動詞派生名詞は，その名の通り動詞から派生した名詞，言

表 1 JEFLL Corpus と Wordbanks online における身体活動を表す動詞派生名詞，動詞及び複合動詞

	JEFLL Corpus				Wordbanks online			
	D-Noun	Verb	CP	Total	D-Noun	Verb	CP	Total
Cry	0	421(100%)	0	421	210(25%)	620(74%)	6(1%)	836
Laugh	2(3%)	74(97%)	0	76	113(14%)	713(85%)	12(1%)	838
Scream	1(5%)	19(95%)	0	20	89(24%)	288(76%)	0	377
Shout	0	102(100%)	0	102	44(12%)	312(88%)	0	356
Smile	20(35%)	35(61%)	2(4%)	57	561(33%)	1,123(66%)	11(1%)	1,695

わば後から出現した名詞であるから動詞に優位性があるのは当然ではあるが，日本人中高校生の動詞に対する極端な偏重は複合動詞を習得する以前に，動詞派生名詞の認識の欠如が疑われる。

一方，英語母語話者では smile は 34%，cry は 26％，scream は 24％，laugh は 15％，shout は 12%（D-Noun と CP との合算）と各動詞派生名詞に一定の使用が認められるが，複合動詞の使用はかなり低い。動詞派生名詞に限定をしても，身体活動を表す動詞派生名詞では英語母語話者でも laugh の 9.6%（総数 125 例中 12 例）を除けば積極的な使用は確認されない。次例参照。

(2) a. They make our smile ..... maybe
　　b. We make smile and my [JP:”Irassiyaimase”].... (JEFLL)

(3) a. Nora gives a stifled cry, runs across the room to the sofa table.
　　b. Finch gave a superior smile.
　　c. He gave a gentle laugh to cover his terror. (Wordbanks online)

3.3　身体動作 : escape, look, stop, visit, walk

表 2 に身体に関わる動作を表す動詞派生名詞，動詞及び複合動詞の頻度を示す。身体動作を表す動詞派生名詞は複数の基本動詞と共起する傾向にある。

日本人中高校生の動詞派生名詞の使用頻度は身体活動と同様に低く，当然複合動詞の使用も著しく低い。一方，英語母語話者では visit の動詞派生名詞と動詞の使用頻度は同じである（D-Noun と CP との合算）。

動詞を含めた総数で検証すると，全体的に複合動詞の使用頻度は英語母語話者でも極めて低いように見えるが，動詞派生名詞に限定をすると，look で 18%（総数 801 例中 144 例），escape で 9%（総数 157 例中 14 例），walk で 7.8％（総数 258 例中 20 例）複合動詞での使用が確認される。名詞単独，あるいは他のコロケーション (e.g., by the look of, an escape of, in a walk, etc.) などの選択肢が

表 2　JEFLL Corpus と Wordbanks online における身体動作を表す動詞派生名詞，動詞及び複合動詞

	JEFLL Corpus				Wordbanks online			
	D-Noun	Verb	CP	Total	D-Noun	Verb	CP	Total
Escape	1(1%)	163(99%)	0	164	143(29%)	333(68%)	14(3%)	490
Look	17(2%)	724(98%)	0	741	657(8%)	7,531(90%)	144(2%)	8,332
Stop	2(2%)	118(98%)	0	120	180(8.3%)	1,983(91.3%)	8(0.4%)	2,171
Visit	3(2%)	127(98%)	0	130	374(48%)	388(50%)	18(2%)	780
Walk	7(1%)	416(95%)	16(4%)	439	238(16%)	1,245(83%)	20(1%)	1,503

ある中で複合動詞という構文に限定をしていることを考慮すれば，決して一概には低いと言えない。特に，look は複合動詞での使用が顕著であると言える。

日本人中高校生はその look でさえ動詞派生名詞での使用は僅か 2%である。無論，look を動詞として使用すること自体に問題はないが，英語母語話者とは明らかに異なる文体で文章を書いていると言える。ただし，walk に限っては，低い頻度ながらも動詞派生名詞より複合動詞が多く観察される。その要因は，take a walk がイディオムとして記憶にしっかりと定着されているためと推察される。次例参照。

(4) a. One day, I was taking a walk.
b. I have taken a walk with my dog. (JEFLL)

(5) a. Let's take a brief look into the four forces, asking three questions.
b. The class decides that someone should make a personal visit to the teacher's family.
c. I wanted to see Mack before he made his nightly escape. (Wordbanks online)

3.4 知的活動 : answer, guess, order, report, talk

表 3 に知的活動を表す動詞派生名詞，動詞及び複合動詞の頻度分析の結果を示す。知的活動を表す動詞派生名詞 guess は，have，make，take と共起し，複合動詞のイディオム的な特徴が表れている。

表 3 JEFLL Corpus と Wordbanks online における知的活動を表す動詞派生名詞，動詞及び複合動詞

	JEFLL Corpus				Wordbanks online			
	D-Noun	Verb	CP	Total	D-Noun	Verb	CP	Total
Answer	45(34%)	85(64%)	2(2%)	132	626(42%)	823(55%)	53(3%)	1,502
Guess	0	23(100%)	0	23	55(7%)	729(91%)	13(2%)	797
Order	16(80%)	4(20%)	0	20	901(66%)	432(32%)	30(2%)	1,363
Note	33(100%)	0	0	33	430(34%)	759(61%)	59(5%)	1,248
Doubt	1(50%)	1(50%)	0	2	234(50%)	185(39%)	52(11%)	471

知的活動を表す動詞派生名詞は，日本人中高校生でも answer，order，note において高頻度の使用が認められる。英語母語話者でも order と doubt では，動詞派生名詞が動詞を凌駕している（D-Noun と CP との合算）。日本人中高校生が示す高い使用頻度は，動詞派生名詞を意識した使用ではなく，すでにこれらの語が上述した smile 同様日本語にカタカナ英語として定着していることに

因ると考えられる。恐らく，note は日本語の「ノート」（ただし，英語の名詞 note と日本語の「帳面」との混同と推察される），order は日本語の「オーダー」，answer は日本語の「アンサー」が日常的に使用されていることで動詞に優先しているのであろう。事実，複合動詞の使用は answer の 2 例を除けば皆無である。

複合動詞に関して，英語母語話者では動詞派生名詞に限定をすると，guess で 19％（総数 68 例中 13 例），doubt で 18.2％（総数 286 例中 52 例），note で 12％（総数 489 例中 59 例）と一定の使用が認められる。特に，guess における複合動詞の顕著な使用頻度は，動詞派生名詞自体 1 例も使用が確認されていない日本人中高校生とは対照的である。単純動詞を好むか，複合動詞を好むかといった表現上の違いは文体上の問題であり文法的な誤りではないので，一見すると見落とされがちであるが，これらの数値から自然な英語表現の実現には一定の割合で複合動詞を使用することが望ましいと言える。次例参照。

(6) a. I don't have the answer of this question, because I have many important item.
b. I don't have this answer. (JEFLL)

(7) a. A person of an obliging disposition gives a peevish answer.
b. He made a note on the list he was holding.
c. Dwight had no doubts about his drinking. (Wordbanks online)

3.5 意志行為 : care, hope, promise, try, wish

表 4 に意志行為を表す動詞派生名詞，動詞及び複合動詞の頻度分析の結果を示す。

表 4 JEFLL Corpus と Wordbanks online における意志行為を表す動詞派生名詞，動詞及び複合動詞

	JEFLL Corpus				Wordbanks online			
	D-Noun	Verb	CP	Total	D-Noun	Verb	CP	Total
Care	7(9%)	39(48%)	35(43%)	81	1,031(53%)	598(31%)	304(16%)	1,933
Hope	18(6%)	267(93%)	2(1%)	287	549(36%)	941(63%)	17(1%)	1,507
Promise	21(81%)	4(15%)	1(4%)	26	194(32%)	397(64%)	26(4%)	617
Try	0	372(100%)	0	372	28(0.8%)	3,747(99.1%)	4(0.1%)	3,779
Wish	6(6%)	93(94%)	0	99	173(19%)	708(78%)	23(3%)	904

意志行為を示す動詞派生名詞は，日本人中高生では promise だけが圧倒的な頻度で名詞として用いられている。一方，英語母語話者には promise にそのよ

うな偏重は観察されず，むしろ動詞の使用が 64% 占めている。また，hope や wish は日本人中高校生では動詞の使用が圧倒的だが，英語母語話者では hope で 37%，wish で 22%（D-Noun と CP との合算）動詞派生名詞での使用が認められる。しかしながら，興味深いことに，care と try には両者の間に共通点が見られる。日本人中高校生，英語母語話者ともに，care はそれぞれ 52%，69%（D-Noun と CP との合算）で動詞派生名詞が確認され，逆に，try はともに動詞の使用が圧倒的である。恐らく，日本人中高校生にとって care は，複合動詞の take care (of) がイディオムとして十分刷り込まれていることで，名詞で使用する際にはこの形式が最初に連想されるのであろう。事実，35 例全て take care (of) であった。ただし，英語母語話者では，take care (of) 以外の複合動詞も確認されている（take care (of) が 287 例，have care が 13 例，give care が 4 例）。

動詞派生名詞に限定をすると，複合動詞に関して，日本人中高校生では 83.3%（総数 42 例中 35 例）で生起する take care (of) を除外すれば，promise と hope に僅かに確認されるに留まるが，英語母語話者では care で 22.8%（総数 1,335 例中 304 例），promise で 11.8%（総数 220 例中 26 例），wish で 11.7%（総数 196 例中 23 例）認められ，名詞単独での使用や他のコロケーション (e.g., in care of, a promise of, a wish list, etc.) などの選択肢を考慮すれば，決して低頻度とは言えない。次例参照。

(8) a. I ‘m taking care of rabbit every day.
b. And he wanted to take care of them.
c. He was so sad and had no hope in his life.(JEFLL)

(9) a. Laura , come here and make a wish on the moon !
b. ...and McKee had a sudden wild hope that he would start it, climb in,...
c. A long time ago I made a promise. (Wordbanks online)

類義関係にある動詞派生名詞は同じ動詞と共起する傾向があることについてはすでに言及したが，hope と wish はともに願望などの意味を有するが，hope は have と wish は make と共起する。同じ願望でもその実現性に伴う動作主の負担に応じて動詞が選択されていると考えられる。負担が軽いものは have と，負担が重いものは make と共起する傾向が観察される。

4．考察

本稿での分析結果から，日本人中高校生の動詞派生名詞と複合動詞の使用頻度の著しい低さは，英語母語話者と比較をすることでより明確になった。その要因として次の 2 つが考えられる。1 つは，動詞と同形の動詞派生名詞に対する認識の低さが挙げられる。いわゆるアウトプットが主となる言語活動では，形式と意味の一致という点において，単純動詞を使用する方が複合的な構造を形成する複合動詞より使用者の負担は，はるかに少ない。ゆえに，まず動詞に意識が行く。加えて，織田（2014: 26）が提唱しているように，日本語は動詞や形容詞を中心に文を組み立てる言語であるが，英語は「名詞に依存する言語」であることを考慮すれば，日本人にはもともと動詞派生名詞は動詞より認識しづらいにことになる。実際，この傾向は本稿で示された日本人中高生に観察された高い動詞志向性と合致する。日本語の干渉も複合動詞の習得を更に困難にしていると考えられる。[4]

もう 1 つは，基本動詞が文中で果たす文法的機能と意味の希薄化に対する知識の欠如が要因と推察される。[5] 複合動詞は主たる意味は動詞派生名詞が担い，基本動詞はいわば文法機能を果たしているにすぎない。それが日本人には理解し難いのである。例えば，「尋ねる」は visit と make a visit，「見る」は look と take a look の 2 つの形式がそれぞれ存在するが，前者は形式と意味との関係が明確であり日本人中高校生から見れば自信を持って使うことが可能だと思われる。しかし，後者では動詞 make の基本的意味である「作る」に引っ張られて「訪問を作る」，あるいは動詞 take の基本的意味である「持っていく」に引きずられて「見ることを持っていく」，というおかしな日本語が連想され日本人中高校生たちに使いにくい印象を与えてしまっているのではないだろうか。堀（2009: 36–37）が指摘しているように，複合動詞の動詞を 1 つの最も基本的意味に限定をしてしまうとその後に出会う「他の意味の理解が困難」になることが考えられる。

複合動詞を使用構文へと昇格させるには，基本動詞の文法的役割と動詞派生名詞の意味的役割を理解させるだけでなく，複合動詞の統語上の柔軟性，対応する単純動詞の不在，韻律上の優位性など使用に伴う労力に値するメリットを生徒に教授することが必要であると考えられる。

どの基本動詞がどの動詞派生名詞と共起するかはイディオムの領域，つまり暗記とされているが，類義関係にある動詞派生名詞は同じ基本動詞を取る傾向

が観察される。例えば，cry，shout，scream，cough など身体の中から出るような negative な行為は give と共起するが，smile，laugh など positive 行為は have と共起する。また，同じ意志行為でも wish，promise は make と，hope は have と共起する。前者は make「何かを作り出す」，つまり動作主にかかる負荷に見合う意志行為と共起し（e.g., make an attempt, make an effort, make a success (of), etc.），後者は have「所有する」，つまり make ほど動作主には負荷がかからない意志行為（e.g., have intention (of), have disregard (for), have a thought (of), etc.）と共起する傾向にあると考えられる。「約束」や「実現可能性が低い願望」より「実現可能性を含意する期待」のほうが動作主にかかる負荷は軽い。

複合動詞を構成する基本動詞は限られているが，動詞派生名詞は数多くある（1.2 節を参照）。このような生産性を享受するには，1 つ核となるプロトタイプ的なパタンの習得が鍵となる（3.1 節の take medicine の例を参照）。それを獲得できれば，類推・推論といった認知機能を働かせながら他の複合動詞を習得することが容易になると思われる。[6]

一方で，動詞派生名詞は複数の基本動詞と複合動詞を形成することは珍しいことではない。例えば，end は make，have，give，take と共起する。これもまた複合動詞の生産的な一面であることを注意しておきたい。

コロケーションは，gradience（段階性）を有しており，kick the bucket から take a look まで統語的にも意味的にもその結束性は段階的である（Quirk et al., 1985: 1162）ことを早期から学習することが望ましい。特に，複合動詞は意味の透明性において句動詞より習得にかかる英語学習者の負担は軽い。

5．結論

本稿では，複合動詞に関して英語母語話者と比較した結果，日本人中高校生の使用頻度の低さ，使用パタンの違い及び動詞派生名詞に対する低い認識が認められた。複合動詞は単純動詞には見られない韻律的・統語的利点を数多く有している。加えて，意味の透明性は習得にかかる負担を軽減させると推察される。複合動詞のこれらの利点を考慮すれば，複合動詞を自在に運用できる能力は英作文に文体的ヴァリエーションを与えるだけでなく英語でのあらゆる言語活動にも有益であると言える。

一方で，複合動詞を含めコロケーションの習得は英語学習者にとってかなり手強い壁である。コロケーションは一見すると構造が単純ゆえに習得が容易で

話法・態・仮定法ほど重要とは思われない。文法は規則があり教えやすいが，コロケーションにもある程度のパタンは見られるものの，その膨大な数とイディオム的意味（構造による）が学習を困難にさせる。英語教師は暗記の前提として複合動詞の構造をしっかりと理解させ，その生産性を享受させるために類推や推論といった作業が暗記学習と等しく言語習得には必要不可欠であることを生徒と共有することが重要である。その認識がパタン学習から創造的学習への転換の鍵となると提唱したい。

本稿で扱えなかった接辞化を伴う動詞派生名詞の検証は今後の課題としたい。

謝　辞

本稿の執筆に際し，大変貴重なご助言とご示唆を賜りました査読委員 3 名の先生方に心より感謝申し上げます。

注

1. JEFLL Corpus では，品詞検索を用いて動詞と名詞を検索することは可能であるが，文法的な間違いをコンピューターが認識できず誤ったタグを付ける可能性がある（投野（編），2007：148–149）。本稿のデータでも名詞と動詞の混同が確認されたため，手動による分析を行った。更に，JEFLL Corpus と Wordbanks online の検索に関して，動詞を指定して品詞検索をかけると現在分詞も過去分詞も当然ヒットするが，この V-ing と V-en 形は，動名詞，形容詞用法も含まれる。形式での検索には限界があり，文の解釈や機能に基づく検索は厳しいため，V-ing と V-en 形は，再度手動での分析により動名詞と形容詞用法は除外した。
2. 学習者コーパスと一般の英語コーパスを比較した研究は，Biber and Reppen（1998），JEFLL Corpus と一般の英語コーパスは，飯尾（2013），内田（2014），Satake（2015）などがある。
3. Wordbanks online のサブコーパスは新聞・パンフレット類・本・雑誌まで多岐に及んでいる。本より雑誌の文体の方が日本人中高校生のレベルにより近いと思われるが，男性誌であることを考慮し，本稿ではアメリカ英語で書かれた本を選択する。
4. Nesselhauf（2000: 231, 239）は，基本動詞の選択の幅と英語学習者の母語が複合動詞の使用を困難にさせていると論じている。実際にドイツ語を母語とする英語学習者はドイツ語の複合動詞の干渉により（e.g., hausaufgaben machen ‘make homework’ in German，but in English ‘do homework’），誤った複合動詞（*make homework）を生成する傾向が報告されている。英語学習者の母語の構造を考慮することも複合動詞の習得には欠かせないと言える（cf. De Cock, 2000: 64–65）。
5. 投野（編）（2007: 40, 67–87）の調査によると，JEFLL Corpus における高頻度動詞の内，基本動詞（do, give, have, make, take）が全て上位 40 にランクインしていたことが示されている。従って，基本動詞の積極的な使用が複合動詞へと発展しない

要因は，基本動詞の文法的機能と動詞派生名詞の認識不足と推察される。

6. Nesselhauf（2003: 239）によれば，類義関係にある名詞ならば常に同じ動詞と共起するとは限らない（e.g., run the risk (of), *run the danger (of), *run the peril (of)）。ゆえに，動詞が自由に生起できないこともきちんと理解しておくことが重要であると指摘している。

参考文献

Aijmer, K. (2001) “Modality in Advanced Swedish Learners’ Written Interlanguage.” In Granger, S, J. Hung and S. Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins, pp. 55–76.

相沢佳子（1999）『英語基本動詞の豊かな世界』開拓社.

秋元実治（編）（1994）『コロケーションとイディオム－その形成と発達－』英潮社.

Biber, D. and R. Reppen (1998) “Comparing Native and Learner Perspectives on English Grammar: A Study of Complement Clauses.” In Granger, S (ed.), *Learner English on Computer*. London: Longman, pp. 145–158.

Brinton, L. J. and M. Akimoto (1999) “Introduction.” In Brinton, L. J. and M. Akimoto (eds.), *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. Amsterdam: John Benjamins, pp. 1–20.

Clardige, C. (2000) *Multi-word Verbs in Early Modern English*. Amsterdam: Rodopi.

De Cock, S. (2000) “Repetitive Phrasal Chunkiness and Advanced EFL Speech and Writing.” In Mair, C and M. Hundt (eds.), *Corpus Linguistics and Linguistic Theory*. Amsterdan & Atlanta: Rodopi, pp. 51–68.

Dixon, R. M. W. (1991) *A New Approach to English Grammar, on Semantic Principles*. Oxford: Oxford University Press.

Hill, J. (2000) “Revising Priorities: from Grammatical Failure to Collocational Success.” In Lewis, M. (ed.), *Teaching Collocation*. Hove: Language Teaching Publications, pp. 47–69.

堀正弘（2009）『英語コロケーション研究入門』研究社.

飯尾豊（2013）「コーパスを用いた日本人学習者の句動詞の使用に関する研究」『熊本大学社会文化研究』第 11 号：35–53.

Koskenniemi, I. (1977) “On the Use of Verbal Phrase of the Type ‘to take revenge’ in English Renaissance Drama.” *Poetica* 7: 80–90.

小屋多恵子（2015）「英語教育とコロケーション」堀正弘（編）『これからのコロケーション研究』ひつじ書房，pp. 23–60.

Koya, T. (2004) “A Comparison of Verb-Noun Collocations in Collected from Revised High School English Textbooks in Japan.”『早稲田大学大学院教育学研究科紀要』11 (2): 55–70.

Lewis, M. (2000)“Learning in the Lexical Approach.” In M. Lewis (ed.), *Teaching Collocation*. Hove: Language Teaching Publications, pp. 155–185.

Live, A. H. (1973)“The *take-have* Phrasal in English.” *Linguistics* 9: 31–50.

文部科学省（2017）『中学校学習指導要領解説　外国語編』
(http://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2017/07/25/1387018_10_1.pdf)
Nation, I. S. P. (2001) *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.
Nesselhauf, N. (2003) “The Use of Collocations by Advanced Learners of English and Some Implications for Teaching.” *Applied Linguistics* 24, 2: 223–242.
Nickel, G. (1968) “Complex Verbal Structures in English,” IRAL 6: 1–21.
Nunburg, G., I. A. Sag and T. Wasow (1994) “Idioms.” *Language* 70: 491–538.
織田哲司（2014）「なぜ now that には that が付いているのか？」『英語教育』63 巻 6 号 : 26.
Quirk, R, S. Greenbaum, G. Leech and J. Svartvik (1985) *A Comprehensive Grammar of the English Language*. London: Longman.
Satake. Y. (2015) “Verb-Noun Collocations and Combinations in the Corpora of Japanese English Learners.”『情報学研究』4 号 : 118–125.　獨協大学情報学研究所.
Sinclair, J. (1991) *Corpus Concordance Collocation*. Oxford: Oxford University Press.
投野由紀夫（編）（2007）『日本人中高生一万人の英語コーパス』小学館.
Traugott, E. (1999) “A Historical Overview of Complex Predicates Types.” In Brinton, L. J. and M. Akimoto (eds.), *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. Amsterdam: John Benjamins, pp. 239–260.
内田富男（2014）「コーパスと英語教育語彙表における基本色彩語の考察―BNC, JEFLL Corpus, CEFR(-J) を用いて」『明星大学研究紀要』第 50 号：19–32.
Wierzbicka, A. (1982) “Why Can You Have A Drink When You Can’t *Have An Eat?” *Language* 58: 753–799.
山本史歩子（2017）「大学生の英作文におけるコロケーション」『青山学院大学　教職研究』第 3 号 : 317–327.

（青山学院大学　Email: syamamoto@ephs.aoyama.ac.jp）

「論文」

A Corpus-Based Study on Japanese EFL Learners' Use of Relative Clause Constructions: CEFR Criterial Feature and Error Analysis

Yuka TAKAHASHI

Abstract

Relative clause (RC) constructions are considered some of the most difficult grammatical items for Japanese EFL (English as a Foreign Language) learners. This study investigates the use of relative clauses by Japanese EFL learners at CEFR, A1 to B2 levels, using L2 learner corpora: the Japanese EFL Learner Corpus (JEFLL: Tono, 2007) and the NICT JLE Corpus (NICT JLE: Izumi, Isahara, & Uchimoto, 2005). The types of RCs were categorized based on the SO Hierarchy Hypothesis (SOHH: Hamilton, 1994) and the frequencies of each RC type were compared against those from a CEFR-based Coursebook Corpus. Error analysis was also conducted for learner corpora. Results show that the frequency order of RCs followed the order predicted by the SOHH at each CEFR level across three corpora and that the frequency increased along the CEFR levels. The error analysis identified various types of structure errors, which are the most frequent error types in both JEFLL and NICT JLE.

1. Introduction

The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR), has influenced the Japanese government's reform plans for English education in Japan. English teachers are encouraged to set learning goals and assess learners' progress using CAN-DO statements, focusing on what the learners can do using English. In related research fields, research programs, such as the English Profile Programme, analyses learner data by extracting so-called criterial features (Hawkins & Filipovic, 2012). According to Hawkins and Filipovic (2012), a criterial feature is a linguistic feature that distinguishes one CEFR level from another. Hawkins

and Buttery (2012) states there are 4 types of criterial features.

- Positive linguistic features (Acquired/ Learnt linguistic features)
- Negative linguistic features (Developing linguistic features)
- Positive usage distributions (Native-like distribution of a correct feature)
- Negative usage distributions (Non-native-like distribution of a correct feature)

A positive linguistic feature is the correct use of a certain linguistic feature, while a negative linguistic feature is its erroneous use. Positive usage distributions are learners' usage distributions similar to those of native speakers, while negative usage distributions are the ones that do not match with native speakers. In this study, the first two (positive linguistic features and negative linguistic features) are focused on by analysing Japanese EFL learners' frequencies of correct use and misuse of RCs.

Hawkins (2009) listed related grammatical items and suggested 20 hypotheses that can be criterial, one of which is the relative clause construction.

If criterial features are extracted from the Japanese EFL learner data, it can be helpful for developing more grammar-specific descriptors for CAN-DO statements at each level. This paper aims to investigate Japanese EFL learners' use of relative clause constructions, using both written and spoken corpora. The CEFR-based Coursebook Corpus is also used in order to compare the input and output based on the CEFR. Furthermore, errors are examined closely.

2. Literature Review

2.1 English relative clauses and their acquisition

Hawkins (2009) suggested in his 20 hypotheses that the use of relative clauses increase along the CEFR levels and that learners use simpler types of RCs, classified by the Noun Phrase Accessibility Hierarchy (NPAH: Keenan & Comrie, 1977)

SU > DO > IO > OBL > GEN > OCOMP

This hierarchy is said to reflect the frequency of the different relative pronoun functions and the presence or absence of specific relative pronoun types in a given

language. Some SLA researchers have claimed that this hierarchical order is applicable for the acquisition of RCs as well. It has been stated that SU is the easiest, followed by DO, IO, OBL, GEN, and OCOMP (Hawkins, 1987; Eckman et al., 1988).

There is other categorization of RCs other than NPAH. The SO Hierarchy Hypothesis (SOHH: Hamilton, 1994) claims that there are four types of RC sentences, depending on the matrix position of the RC and the depth of embedding. S means Subject, and O is Object. To illustrate the four types of RCs, Hamilton (1994) showed the following example sentences:

OS They saw **the boy who** entered the room.
OO A man bought **the clock that** the woman wanted.
SS **The man who** needed a job helped the woman.
SO **The dog that** the woman owns bit the cat.

OS < OO/SS < SO
< = is implicated by

Izumi (2003) highlighted that both the matrix position of RC and depth of embedding affect ease or difficulty of processing RCs. It is suggested that OS is the easiest type followed by OO and SS, and the most difficult type is SO. The easiest type of RC modifies the object of the matrix sentence and relative pronoun functions as the subject of the RC. In contrast, the most difficult type of relative clause modifies the subject, and the relative pronoun functions as the object.

Takahashi (2016) analysed the use of RCs in JEFLL and Coursebook Corpus. She used NPAH (Keenan & Comrie, 1977) to categorize RCs. Her study showed that the frequency itself could be a good indicator of proficiency level as the use rate per person increased along the CEFR level (A1 to B2). Also, the order of NPAH was supported as the frequency increased along the level while keeping NPAH order at each level.

Avoidance (Schachter, 1974) of the use of RCs may be possible, since they are some of the most difficult grammar items for Japanese EFL learners. Murakoshi (2015) analysed 209 1st to 3rd grade Japanese High School students' English writing in order to reveal usage frequencies of various grammatical items. For the usage of RCs, he stated that 85% of the 3rd year high school students do not use RCs even though they are

familiarized with RCs from junior high school and continuously practiced using them in high school English classes. However, since the task was spontaneous writing, it was not clear whether they could produce them if they were told to, or whether they had not yet acquired them.

Biber et al. (1999) illustrated the RC distributions in different registers. They revealed the frequency of each relativizer in the *Longman Grammar of Spoken and Written English Corpus* (LGSWE), which comprises 40 million words of texts in American and British English. The corpus is analysed from the perspectives of four registers: conversation, fiction, news, and academic prose. Overall, RC constructions occur far more frequently in academic prose.

2.2 Japanese relative clauses

Studies comparing Japanese and English relative clauses have provided useful insights when analysing Japanese EFL learners' relative clause errors in relation to L1 transfer. Some researchers have argued that some Asian languages do not share the same properties as English and European languages. Matsumoto (1988), for instance, mentioned that in Japanese, there are no relative pronouns, and Japanese has "gapless" relative clause-like constructions and "gapped" relatives. "Gapped" and "gapless" relative clause-like constructions are shown below (Matsumoto, 1988: 167).

(1) "Gapped"
[[hon-o kata] gakusei] – wa doko desu ka.
Book –ACC bought student TOP where is QUES.PART.
'Where is the student who bought a book?'
(2) "Gapless"
[[Atama-ga yokunaru] hon]
head-NOM improve book
'the book (by reading) which (one's) head improves'

Matsumoto also stated that the Japanese appositive clause construction has the same surface structure as the typical RC constructions. Based on these, Comrie (1998:51) proposed that the Japanese relative clause-like constructions should be called, "attributive clauses".

In addition, Ozeki (2011) highlighted that Japanese children can produce surprisingly complex relative clause sentences as below at a very early age around two years old:

[[kureyon moratta] oniityan-ni moratta] ametyan
'the candy [which I got from the guy [from whom I got crayons]] (Sumi 2: 10)

Examples from Ozeki (2011: 180)

Ozeki (2011) argued that this kind of complex structure is possible since there are no formal grammatical restrictions between the head noun and its relative clause in Japanese. Therefore, in the case of Japanese, one can attach simple sentences to the head nouns without grammatical restrictions.

These differences between English and Japanese language structure are important when analysing RC errors produced by Japanese EFL learners. Learners may apply the knowledge of flexible RC structure in Japanese when they produce English RC sentences, even though English RC structures are strictly ruled by its grammar.

2.3 Error analysis

In order to avoid comparative fallacy (Bley-Vroman, 1983) in data analysis, Lüdeling and Hirschmann (2015) suggested the following ways of detailed error analysis. Learner Utterance (LU) is utterance produced by learners, and target hypothesis (TH) is the correct usage. There might be multiple THs if there is more than one way to correct the LU.

LU: She must saved money.
TH1: She must have saved money.
TH2: She must save money.

(Lüdeling & Hirschmann, 2015)

In the detailed error analysis of the structure errors in this study, this way of annotation is used in order to find out possible causes of errors.

In the review, two types of categorizations of RC, NPAH and SOHH, are

explained. In this study, SOHH, which includes both the matrix position of RC and depth of embedding, will be used for the RC categorization and analysis.

3. Research Questions and Method

3.1 Research Questions

Studies related to relative clauses in SLA have mainly used elicitation task, which forces learners to use relative clauses. However, corpus data can reveal spontaneous use of relative clause as criterial features. Furthermore, corpus data specific to Japanese EFL learners needs to be focused as a useful means to gain insights for errors that might be characteristic to Japanese learners. Considering the above, three research questions are set as follows:

RQ1: Do the frequencies of RCs increase along the levels? Are they useful for distinguishing one CEFR level from another as a criterial feature?
RQ2: Does the order of difficulties predicted by SOHH correspond to the use of RCs across the CEFR levels?
RQ3: Which types of errors frequently occur in JEFLL and NICT JLE? If subcategorization of the structure error (STR) is possible, what kinds of errors are included in STR?

3.2 Relative pronouns examined in this study

This paper analyses RC constructions involving *that, which, who, whose* and *whom.* Zero-relatives are not included in this analysis.

3.3 Corpora

The following are brief descriptions concerning the corpora consulted in this study.

3.3.1 The JEFLL Corpus

The JEFLL (Japanese EFL Learner) Corpus is a written learner corpus (Tono, 2007). It consists of 10,063 compositions written by Japanese junior and senior high school students. The total number of words is 669,281 running words. Participants were asked to choose one out of six topics to write about in 20 minutes without any

preparation time. The use of dictionaries was not allowed. Topics include: a) Urashima, b) Rice or Bread, c) Festivals; d) Earthquake, e) Otoshidama, and f) Bad Dreams. For this study, a version of the JEFLL data reclassified by the CEFR levels was used. This version of data was developed by the KAKEN project, which was led by Tono (2016).

3.3.2 The NICT JLE Corpus

The NICT JLE Corpus is a spoken corpus, based on the Standard Speaking Test (SST) developed by ALC Press. The SST is a 15-minute face-to-face oral interview, consisted of the following five stages: 1) warm-up questions (3−4 minutes), 2) single picture description (2−3 minutes), 3) role play with the interviewer (1−4 minutes), 4) picture sequences (2−3 minutes), and 5) wind-down questions (1−2 minutes). The performance in the interview is assessed by two trained raters and is classified into one of the nine SST levels. Most participants are adults (Izumi, Isahara & Uchimoto, 2005).

3.3.3 The Coursebook Corpus

The CEFR-based Coursebook Corpus (Coursebook Corpus, hereafter) includes texts from 105 coursebooks published in Europe, which were all written according to the CEFR design specifications (1,761,520 running words). This corpus was also developed as part of the KAKEN project directed by Tono (2016).

3.4 Extraction of RCs

All the sentences containing relative pronouns were extracted from each corpus using AntConc (Anthony, 2014) and the Sketch Engine (Kilgarriff et al., 2014). Each surface form of the relative pronouns *that, which, who, whose* and *whom* was directly inserted into the search field to extract all the examples. Sentences without relative pronouns were manually removed from the list.

Table 1 below shows the total number of relative clause sentences at each level in three corpora. Analyses were conducted on 1807 examples from JEFLL, 1361 instances from NICT JLE, and 4704 instances from Coursebook Corpus.

Table 1: Total number of relative clause sentences in each corpus

Corpus	A1	A2	B1	B2	C1	C2	Total
JEFLL	129	861	794	23	-	-	1807
NICT JLE	38	528	605	190	-	-	1361
Coursebook Corpus	49	356	1199	1958	1034	108	4704

3.5 Annotations

Each sentence was annotated manually for the different types of relative pronouns (surface form) and constructions of relative clauses (SOHH) using Microsoft Excel spreadsheets. Tags are summarized below.

a) Surface form (*that, which, who, whose, whom*)
b) SOHH types (OS, OO, SS, SO)

3.6 Error types

For the learner data, error tags were also added. Error categories were defined based on Takahashi (2016), which includes the following seven types. Possible correct answers are answers requiring minimum correction.

a) Missing antecedents (MAT)
   At that time, <MAT> *who eat breakfast with me is my mother.*
b) Missing prepositions (MPR)
   He went to the place which he used to live <MPR>.
c) Missing relative subjects (MRS)
   I can meet my friends who <MRS> *haven't seen a long time.*
d) Missing relative objects (MRO)
   Our play was popular among the people who came and see <MRO>.
e) Resumptive pronouns (RSP)
   I take out the thing which <RSP>***it***</RSP> *is important to me.*
f) Wrong selection (SEL)
   Our class had a drama <SEL>*which*</SEL> *name is "Unexpected Guest."*
g) Ungrammatical structures in relative clauses (STR)
   We sang a song which is a famous singer.
   He made a magazine which called "Love from ryugujo."

Takahashi (2016) argued that subcategorization of *Ungrammatical structures in relative clause* (STR) needs to be done, since STR appears to be the most frequent type, which contains various structural errors, whereas the others are more specific to a particular element of the RC. Therefore, in order to provide a breakdown of STR, a detailed error analysis is conducted by examining each instance closely and giving each instance a

possible TH (Lüdeling & Hirschmann, 2015). Those errors are manually annotated for the following: 1) TH, 2) error information about surface structure, and 3) possible reasons why the error occur.

4. Results

4.1 RC frequencies

Table 2 and Table 3 show the usage rate of overall RCs across the CEFR levels in JEFLL and NICT JLE. In each table, it can be seen that the usage rate per person gradually increases along the CEFR level. This indicates that regardless of whether the RC sentences contain errors, the frequency of spontaneous RC productions can be a good indicator of their proficiency level as a criterial feature. Comparing two corpora, the usage rate per person in NICT JLE, which comprises spoken data, showed a relatively higher rate than JEFLL across the CEFR levels.

Table 2: The usage rate of overall RCs across the CEFR levels in JEFLL

JEFLL	A1	A2	B1	B2
Number of files (*n*)	3507	4956	1529	46
RC sentences	129	861	794	23
Usage rate per person	0.04	0.17	0.52	0.50

Table 3: The usage rate of overall RCs across the CEFR levels in NICT JLE

NICT JLE	A1	A2	B1	B2
Number of files (*n*)	260	718	263	40
RC sentences	38	528	605	190
Usage rate per person	0.15	0.74	2.30	4.75

4.2 SOHH type frequencies

Based on the raw frequencies and the size of the sub-corpora, normalized frequencies of SOHH in every 100,000 words are shown in Table 4.

Table 4 illustrates that the frequency order largely followed the order predicted by SOHH in each level, with partial differences. The frequency order follows OS > OO > SS > SO in JEFLL, whereas in NICT JLE and Coursebook, the order was OS > SS >

Table 4: Normalized frequencies of SOHH types across the CEFR levels (per 100,000 words)

Corpus	SOHH	A1	A2	B1	B2	C1	C2	Total
JEFLL	OS	41	137	187	99	-	-	464
	OO	31	78	105	77	-	-	291
	SS	14	42	53	66	-	-	175
	SO	6	14	21	11	-	-	52
	Unknown	2	0	1	0	-	-	3
	Total	94	271	367	253	-	-	985
NICT JLE	OS	13	50	112	179	-	-	354
	OO	11	14	30	57	-	-	112
	SS	11	31	57	89	-	-	188
	SO	2	4	8	11	-	-	25
	Unknown	0	3	1	4	-	-	8
	Total	37	102	208	340	-	-	682
Coursebook	OS	24	75	167	222	242	253	983
	OO	6	18	39	44	69	60	236
	SS	13	23	40	66	65	49	256
	SO	0	4	5	8	9	18	44
	Total	43	120	251	340	385	380	1519

OO > SO. More object relatives at subject position are used in JEFLL, and more subject relatives at subject position are used in NICT JLE and Coursebook Corpus.

In order to statistically test the frequencies across the CEFR levels, chi-square test and residual analysis were carried out. Tables 5, 6, and 7 below indicate the standard residuals in three corpora (SOHH types, CEFR level)

In Table 5 (JEFLL), the overall chi-square test was statistically significant (χ^2 (15) = 245.763, p<.01, V=0.01). The frequencies of all the SOHH types (OS, OO, SS, and SO) at A levels were found to be significantly lower than expected; however, they showed the highest frequencies at the B1 level. The SOHH frequencies seem to be a good indicator to differentiate between A1 and B1 levels.

In Table 6 (NICT JLE), the overall chi-square test was statistically significant (χ^2 (15) = 413.572, p<.01, V=0.011). NICT JLE shows a clear division between A-levels and B-levels in terms of the significantly lower frequencies of OS, OO, and SS at A-levels compared against significantly higher frequencies at B-levels. It is noteworthy that the most difficult SO type increased at B1 level in NICT JLE.

Table 5: The results of chi-squared test and residual analysis (JEFLL)

SOHH	A1	A2	B1	B2
OS	-10.474**	0.819	8.361**	-0.883
OO	-6.919**	0.509	5.42**	-0.024
SS	-5.409**	0.718	3.58**	1.261
SO	-2.968**	-0.325	2.971**	-0.271
Unknown	1.493	-1.711	0.62	-0.308
Other words	13.848**	-0.951	-10.966**	0.231

χ^2 (15) = 245.763, p<.01, Cramer's V = 0.010, *p<.05, **p<.01

Table 6: The results of chi-squared test and residual analysis (NICT JLE)

SOHH	A1	A2	B1	B2
OS	-7.458**	-8.697**	9.537**	9.695**
OO	-2.422*	-5.239**	4.208**	6.157**
SS	-4.998**	-4.862**	5.578**	6.074**
SO	-1.535	-1.918	2.161*	1.896
Unknown	-1.625	2.66**	-2.14*	0.672
Other words	9.452**	10.883**	-11.608**	-13.019**

χ^2 (15) = 413.572, p<.01, Cramer's V = 0.011, *p<.05, **p<.01

Table 7: The results of chi-squared test and residual analysis (Coursebook)

SOHH	A1	A2	B1	B2	C1	C2
OS	-12.353**	-14.208**	-1.202	10.92**	9.304**	3.26**
OO	-5.8**	-6.667**	-0.274	1.789	8.308**	1.691
SS	-5.418**	-6.895**	-2.96**	7.744**	4.39**	0.105
SO	-2.275**	-1.954	-1.031	1.625	2.188*	2.435*
Other words	14.906**	17.244**	2.486*	-13.018**	12.902**	-3.697**

χ^2 (20) = 761.123, p<.01, Cramer's V = 0.010, *p<.05, **p<.01

In Table 7 (Coursebook), the overall chi-square test was statistically significant (χ^2 (20) = 761.123, p<.01, V=0.01). Proportional use of SOHH types between the lower three groups (A and B1 levels) and the upper three groups (B2 and C levels) were seen. At A levels, all the four types of relative clauses were lower than expected, which was all statistically significant, except for SO at A2. At B1, all the four types became fairly frequent, and no statistical difference in observed frequencies was found against expected frequencies. At B2, however, OS and SS became more frequent than

expected, and all the four types were found to be significantly more frequent than expected at B2 level. OO and SS are considered to be equally difficult, positioned in the middle, according to the SO hierarchy, but as far as Coursebook Corpus is concerned, SS seemed to be more widely used at the intermediate levels than OO.

The results of chi-square tests and residual analysis show that the frequencies in each of the three classifications drew a clear line, especially between the groups lower than B1 and those above B1. Therefore, each corpus seemed to show a common cut-off point in frequencies to distinguish the upper CEFR levels from the lower ones. This indicates that SOHH type frequencies in all corpora at each level followed SOHH order with partial differences, and their frequency increased along the level, thereby keeping the SOHH frequency order.

4.3 Error Analysis

Tables 8 and 9 indicate the frequencies and percentages of correct use and misuse of relative clauses in two learner corpora. It was found that the RC error rate in JEFLL and NICT JLE was 22.47%, and 12.65%, respectively. The majority of RCs were used correctly.

Table 8: Frequencies and percentages of correct use and misuse of RCs (JEFLL)

JEFLL	RC sentences	%
Correct use	1401	77.53%
Errors	406	22.47%
Total RC sentences	1807	100.00%

Table 9: Frequencies and percentages of correct use and misuse of RCs (NICT JLE)

NICT JLE	RC sentences	%
Correct use	1189	87.35%
Errors	172	12.65%
Total RC sentences	1361	100.00%

Tables 10 and 11 below summarize the frequencies and percentages of correct use and misuse across the CEFR levels. To make the two corpora comparable, 200 samples were randomly sampled from each corpus.

Table 10: Frequencies of RC errors across the CEFR levels
(Random sampling of 200 cases per level: JEFLL)

Error types	A1		A2		B1		B2		Total	
MAT	1	0.5%	3	1.5%	3	1.5%	0	0.0%	7	0.9%
MPR	6	3.0%	0	0.0%	3	1.5%	16	8.0%	25	3.1%
MRO	0	0.0%	1	0.5%	0	0.0%	0	0.0%	1	0.1%
MRS	6	3.0%	2	1.0%	3	1.5%	0	0.0%	11	1.4%
RSP	4	2.0%	2	1.0%	3	1.5%	0	0.0%	9	1.1%
SEL	14	7.0%	9	4.5%	7	3.5%	0	0.0%	30	3.8%
STR	30	15.0%	19	9.5%	27	13.5%	8	4.0%	84	10.5%
Error total	61	30.5%	36	18.0%	46	23.0%	24	12.0%	167	20.9%
Correct use	139	69.5%	164	82.0%	154	77.0%	176	88.0%	633	79.1%
Total	200	100.0%	200	100.0%	200	100.0%	200	100.0%	800	100.0%

Note: Bootstrap sample was used for A1, due to its small sample size.

Table 11: Frequencies of RC errors across the CEFR levels
(Random sampling of 200 cases per level: NICT JLE)

Error types	A1		A2		B1		B2		Total	
MAT	1	0.5%	1	0.5%	2	1.0%	0	0.0%	4	0.5%
MPR	2	1.0%	2	1.0%	1	0.5%	2	1.0%	7	0.9%
MRO	1	0.5%	0	0.0%	0	0.0%	0	0.0%	1	0.1%
RSP	5	2.5%	2	1.0%	3	1.5%	6	3.0%	16	2.0%
SEL	1	0.5%	11	5.5%	0	0.0%	0	0.0%	12	1.5%
STR	36	18.0%	25	12.5%	8	4.0%	11	5.5%	80	10.0%
Error total	46	23.0%	41	20.5%	14	7.0%	19	9.5%	120	15.0%
Correct use	154	77.0%	159	79.5%	186	93.0%	181	90.5%	680	85.0%
Total	200	100.0%	200	100.0%	200	100.0%	200	100.0%	800	100.0%

Looking at the proportion of each error more closely, structure error (STR) is the most frequent, and it is more than a half of the total errors in JEFLL (10.5% out of 20.9%) and NICT JLE (10.0% out of 15.0%). As a common tendency in two corpora, STR frequencies are lower at B levels than A levels, but they remain the most frequent error type, even at B2 level. Because of this occurrence, more detailed error analysis was carried out.

Other than STR, in JEFLL, it should be noted that selection error (SEL) gradually declined along the level, while missing preposition error (MPR) increased, especially at B2. This illustrates the occurrence of selection error in the RC productions of low

level learners, but as they start to use more complex structures involving prepositions, they start dropping the prepositions. This indicates that some of the frequent errors for low proficient and more proficient learners can be different and that making errors is not necessarily negative since the learners started using more complex structures.

4.4 Detailed Error Analysis

In 4.3, the most frequent error was found to be the structure error (STR). However, since many types of errors seemed to be mixed in this category, subcategorization of STR is attempted based on the surface structure. Tables 12 and 13 present the results of subcategorization in STR and the frequencies of the subcategories.

Table 12: STR error frequencies with subcategories (JEFLL: 217 instances)

Subcategories of STR	Frequencies	%
Structure/word order errors	75	34.56
RP + be-verb errors	19	8.76
Missing be-verb	76	35.02
Unnecessary (direct translation from Japanese)	22	10.14
Use of Japanese	16	7.37
Incomplete	9	4.15
Total	217	100.00

Table 13: STR error frequencies with subcategories (NICT JLE: 105 instances)

Subcategories of STR	Frequencies	%
Structure/word order errors	43	40.95
RP + be-verb errors	21	20.00
Missing be-verb	18	17.15
Unnecessary (direct translation from Japanese)	17	16.19
Incomplete	6	5.71
Total	105	100.00

The error types in tables 12 and 13 are explained with examples below.

- ***Structure/word order errors***
  Errors include complex grammatical errors and word order errors. This comprises errors that are difficult to state the common cause of the errors.

e.g.) *adults who are supposed to send children who and themselves know each other* [JEFLL, A2]

- ***RP (relative pronoun) + be-verb errors***
  Learners may understand that RC is one of the post-nominal modifiers, but the sentence structures following *RP + be* is incorrect. Learners may consider *RP + be* as a fixed phrase.

e.g.) *So they watched cinema which is trouble of plane* [NICT JLE, A1]

- ***Missing be-verb in relative clause***
  Be-verb following RP is dropped.

e.g.) *I like bread which ● made by us.* [JEFLL, B1]

- ***Unnecessary* (direct translation from Japanese)**
  This type is not necessarily errors but unnatural use of relative clauses, which may occur due to direct translations from Japanese phrases. The underlined part in the example below might be a direct translation from *otona no hito* 'adults.'

e.g.) *<jp>Otoshidama</jp> is some money which people who are adults give children.* [JEFLL, A2]

- ***Incomplete***
  Incomplete RC sentences are errors because they drop a necessary part of the RC sentence.

e.g.) *The man who is running on the road ●.* [NICT JLE, A1]
Suddenly a man who had a knife and gun ●. [JEFLL, A2]

The most frequent type was *structure/word order errors* in both corpora. It was not possible to identify the common causes of errors for this type, but all included

complex structural errors.

RP + be-verb errors can be considered as a part of *structure/word order errors*, but it has a common tendency in that the learners use RP and be-verb as a set. This might occur when learners gain input of OS-type RCs, which is found to be most frequent and easy, and assume that the be-verb comes right after the RP all the time. This indicates that learners know they can modify nouns using RPs, but they make structural errors when constructing whole RC sentences.

Missing be-verb error was frequently observed especially in JEFLL. Looking at the surface form, they are just missing be-verb; however, there might be some causes related to the use of RCs. The typical example of a missing be-verb (LU), the possible correct answer (TH), and possible causes of errors a) ~ c) are shown below:

LU: *I like bread which made by us.* [JEFLL, B1]

a) dropping be-verb of passive in RC.
   TH: I like bread which <u>is</u> made by us.

b) the grammar of RC construction and post nominal participle construction are confused
   TH: I like bread made by us.

c) OO type (which is more difficult than OS type) has not been acquired.
   TH: I like bread (which) <u>we</u> made.

The first possible answer (TH) for the LU is *I like bred which <u>is</u> made by us*, and this is based on the explanation that LU is dropping the be-verb of passive in RC. Inserting *is* makes the sentence correct. On the other hand, as the second TH shows, there is a possibility that the learner is using relative *which* and post nominal participle construction at the same time. For this case, removing *which* makes the sentence correct (*I like bread made by us*.). The last TH shows the possibility that the learners have not acquired OO type which is more difficult than OS type. To make this sentence correct from this perspective, *we* is inserted after the relative pronoun *which*.

Moreover, there is a possibility that learners make such errors when they use passives or post-nominal participle constructions within RCs, while they can correctly use them outside the RC sentences. In the entire JEFLL data, six learners used both RCs and passives or post-nominal participle constructions in their writing. Example

sentences from each learner are shown below.

Table 14: Example sentences produced by six learners who may be confused with the use of relative clauses, passives, and post-nominal participle constructions (JEFLL).

Learner 1	✓	…a bird called Tsuru…	PNP
	×	…a Omiya which built at Tango…	RC + passive, or PNP
Learner 2	✓	The ship was named "dream".	Passive
	×	…a ship that made from woods.	RC + passive, or PNP
Learner 3	✓	I was maked (made)…	Passive
	×	…a video that called about school festival.	RC + passive, or PNP
Learner 4	✓	…the special stage are built by senior students.	Passive
	×	…the place which we are called "Stage"	RC + passive, or PNP, OO type construction
Learner 5	✓	…pretty thing which was made by…	RC + passive
	×	…buresuretto which made of bi-zu.	RC+ passive, or PNP
Learner 6	×	…bread that made by XX.	RC + passive, or PNP
	×	…bread that made by XX.	RC + passive, or PNP

Note: Relative Clause (RC), Post-Nominal Participle Construction (PNP)

Learners 1, 2, 3, 4 used passives and post-nominal participle constructions correctly; however, they made errors when those are used with relative pronouns. This indicates that: 1) they could not construct passives in RCs and 2) they confused the use of RCs and post-nominal participle constructions. Learner 6 used *RP + made by* twice and both examples dropped the be-verb. This might be because Learner 6 used *made by* as a fixed phrase. On the other hand, Learner 5 used *RP + made by* correct, but *RP + made of* wrong. Learners 5 and 6 were both at B1 level, and this shows the feature of middle level learners' interlanguage, at which the knowledge of the form is not fully stabilized, thus producing occasionally ill-formed sentences. For those errors, it is also possible to say that since they have not acquired OO (or SO) type structure, they could not construct the latter part of the relative clause sentence after a relative pronoun. The input and output of OO (or SO) type may also be important for learners to be able to express what they want to say.

5. Discussion

In this section, each research question is revisited, and the study's results and implications will be discussed.

RQ1: Do the frequencies of RCs increase along the levels? Are they useful for distinguishing one CEFR level from another as a criterial feature?

The results showed that the RC frequencies increased along the level, and the ability to produce RC sentences spontaneously can serve as a positive linguistic features distinguishing A1 level from B1 level learners, which confirms Hawkins's hypothesis (2004). As learners become more proficient, more complex structures are frequently used. It is said that the RCs are used more frequently in writing than in speaking, yet NICT JLE showed more frequent RCs than JEFLL. One probable reason is because the Japanese EFL learners included in NICT JLE were adults, who could afford to pay examination fees, whereas JEFLL comprises written production of Japanese junior high and high school students. Moreover, considering the EFL learning environment in Japan, which focuses more on writing than speaking, one student's level of performance can be different for writing and speaking. Since the learners in the two corpora are not identical, A2 learners in NICT JLE may perform better and produce more RC sentences than do A2 learners in JEFLL, which also might mean that A2 learners in NICT JLE may perform better in writing than A2 learners in JEFLL. Moreover, in the Standard Speaking Test include picture description tasks, which force learners to describe particular people or things in detail, which may have led relatively short and simple but frequent production of RC sentences. Further, the interviews may have helped learners to produce a greater number of utterances. In JEFLL, what the students tried to write based on their experience might have been more complex and original.

RQ2: Does the order of difficulties predicted by SOHH correspond to the use of RCs across the CEFR levels?

It was found that the frequencies of RCs followed the order predicted by SOHH in all corpora, with partial differences. Hamilton's SOHH order shows that the difficulty of OO and SS is almost similar. JEFLL showed the frequency order OS > *OO*

> *SS* > SO, whereas NICT JLE and Coursebook Corpus demonstrated the OS > *SS* > *OO* > SO order at almost all levels. It should be noted that the difference between OO and SS might have occurred due to task effects in NICT JLE and coursebooks. For JEFLL, spontaneous production of free writing may have encouraged learners to write a) who does what to whom and b) why some particular things or people are important or their favourite. Such conditions may have allowed learners to use RC sentences with objects. Furthermore, the stories are already in the students' minds, such as their actual experiences or imaginations, which allows learners to access various complex stories instantly. On the other hand, tasks in NICT JLE include picture descriptions and picture sequences, which prompt speakers to explain particular things in the picture, such as how they look like and where they are, which may have elicited more subject RCs. It might be difficult to create picture description tasks that elicit various grammatical items and object relatives at the same time by using only few pictures. Coursebooks also have similar attributes as NICT JLE. There are limited spaces for pictures, reading materials, and exercises. Coursebooks are based on oral communication, which may focus mainly on simple OS type or subject RCs.

RQ3: Which types of errors frequently occur in JEFLL and NICT JLE? If subcategorization of the structure error (STR) is possible, what kinds of errors are included in STR?

The RC error rates in JEFLL and NICT JLE were not very high, and remained at 12% and 22%, respectively. Avoidance (Schachter, 1974) of the use of relative clauses may have occurred when students avoided using complex structures as they were afraid of making mistakes. Even though it was difficult to determine errors as being negative linguistic features due to the small sample size, it was found that there were different causes of errors in the most frequent error type, structure error (STR). A detailed error analysis and subcategorization of STR revealed that the STR included errors involving some grammatical items that might be confusing for learners to distinguish. RCs, passives, and other post-nominal modifications are introduced separately in different sections in the English textbooks. Considering these facts, after their initial introduction, those grammatical items need to be explained repeatedly concerning a) their usage together (e.g. passives in RCs), b) which part of the grammar items is similar and which part is different (e.g. differences between RCs and other post-

nominal modifications), or c) how they are different from Japanese, as English RCs are governed by strict rules, whereas Japanese is grammatically more flexible (Ozeki, 2011). There is a common tendency in that learners try to modify nouns, using relative pronouns by attaching simple sentences to the head noun, which results in structure errors.

6. Conclusion

This study aimed to examine Japanese EFL learners' use of relative clauses as criterial features. Written and spoken Japanese EFL learner corpora were analysed in order to extract criterial features, and a CEFR-based Coursebook Corpus was also analysed for comparison. Results show that a) spontaneous use of relative clauses and its frequency can serve as criterial features; b) the SOHH frequency followed the hierarchy predicted by SOHH at each level; and c) the overall frequencies increased along the level. Error analysis revealed that the most frequent error type, which contains structural errors (STR), can be subcategorized, and similar grammatical items need to be continuously taught.

There are some methodological limitations in this study. First, zero-relatives need to be extracted for analysis using regular expressions in order to more accurately capture the use of relative clauses. Second, error tagging should be done by more than one annotator in order to gain reliability in the tagging. Finally, Japanese English textbooks need to be added to the dataset, in order to investigate the relationship between input and output.

The information of RC type frequencies and error type frequencies at each CEFR level, may contribute to provide specific descriptions for each level, based on the data pertaining to Japanese EFL learners. Using relative clauses is especially important when reaching B level from A level, so that knowing frequently used RC types and frequently committed errors by B level learners may help teachers, learners, and teaching material developers by providing them with clearer objectives.

References

Anthony, L. (2014) AntConc (Version 3.4.3) [Computer Software]. URL: http://www.laurenceanthony.net/

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. and Quirk, R (1999) *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Ltd.

Bley-Vroman, R. (1983) "The comparative fallacy in interlanguage studies: The case of systematicity." *Language Learning 33*, 1: 1–17.

Comrie, B. (1998) "Attributive clauses in Asian languages: Towards an areal typology." In W. Boeder, C. Schroeder, K. Wagner, and W. Wildgen (eds.), *Sprache in Raum und Zeit: In memoriam Johannes Bechert*. Tubingen: Gunter Narr, pp. 51–60.

Eckman, F. R., Bell, L. and Nelson, D. (1988) "On the generalization of relative clause instruction in the acquisition of English as a second language." *Applied Linguistics 9*, 1: 1–20.

Hamilton, R. (1994) "Is implicational generalization unidirectional and maximal? Evidence from relativization instruction in a second language." *Language Learning 44*, 1: 123–157

Hawkins, J. A. (1987) "Implicational universals as predictors of language acquisition." *Linguistics 25*, 3: 453–474.

Hawkins, J. A. (2009) "*Cambridge / UCLES-RCEAL research projects*". Internal report.

Hawkins, J., and Buttery, P. (2010) "Criterial features in learner corpora: Theory and illustrations. *English Profile Journal 1*, 1: 1–123.

Hawkins, J. A. and Filipović, L. (2012) *Criterial Features in L2 English, Specifying the Reference Levels of the Common European Framework*. Cambridge: Cambridge University Press.

Izumi, S. (2003) "Processing difficulty in comprehension and production of relative clause by learners of English as a second language." *Language Learning 53*, 2: 285–323.

Izumi, E., Isahara, H. and Uchimoto, K. (eds.), (2005) *Nihonjin 1200 nin no eigo supikingu kopasu*. [*English speaking corpus of 1200 Japanese EFL learners*]. Tokyo: ALC Press.

Keenan, E. L. and Comrie, B. (1977) "Noun phrase accessibility and universal grammar." *Linguistic Inquiry 8*, 1: 63–99.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel., V. (2014) The Sketch Engine: ten years on. *Lexicography 1*, 1: 7–30.

Lüdeling, A. and Hirschmann, H. (2015) "Error annotation systems." In Granger, S., Gilquin, G., and Meunier, F. (eds.), *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, pp. 135–158

Matsumoto, Y. (1988) "Semantic and pragmatics of noun-modifying constructions in Japanese." *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society 14*, 166–175.

Murakoshi, R. (2015) "The Development of Syntactic Complexity in Japanese High School Students' English Compositions." *ARCLE Review 9*, 17–26.

Ozeki, H. (2011) "The acquisition of relative clauses in Japanese". In Kidd, E. (ed.), *The acquisition of relative clauses: Processing, typology and function*. Amsterdam: John Benjamin Publishing Company, pp. 173–194.

Schachter, J. (1974) "An error in error analysis." *Language Learning 24*, 2: 205–214.

Takahashi, Y. (2016) “Relative Clause Constructions as Criterial Features: A Corpus-based Study.” *Selected Papers from 2016 PAC/The twenty-fifth International Symposium on English Teaching*, 236–249.

Tono, Y. (ed.) (2007) *Nihonjin chuko-sei ichimannin no eigo corpus* [*The JEFLL Corpus: A corpus of 10,000 English Essays by Japanese Secondary School Students*]. Tokyo: Shogakukan.

Tono, Y. (ed.) (2016) *The CEFR-J RLD project: Developing Grammar, Text and Error Profiles Using Textbook & Learner Corpora.* The Final Report of the Grant-in-Aid for Scientific Research (A) (no. 24242017). Unpublished internal report. Retrieved from http://cefr-j.org/PDF/TonoKaken2012–2015FinalReport.pdf on 31 January, 2018.

（高橋　有加　広島大学外国語教育研究センター　Email: y-takahashi@hiroshima-u.ac.jp）

「論文」

How Have Political Interests of U.S. Presidents Changed?: A Diachronic Investigation of the State of the Union Addresses through Topic Modeling

Naoki KIYAMA

Abstract

In this study, I demonstrate that the topic modeling method, more specifically the latent Dirichlet allocation (LDA), is a useful method to investigate how U.S. presidents' political concerns have changed since the foundation of the United States. By applying the LDA to State of the Union Addresses, I obtain four main topics ((i) internal issues related to the federal government; (ii) international affairs tied with territorial disputes; (iii) worldwide warfare; and (iv) social welfare) and argue that each topic is a reflection of its historical background. Last, I proposed that a transition in the key topics occurred before and after World War II. In other words, the Presidents' main political concerns were influenced by whether the United States had attained the status as the world leader.

1. Introduction

There are some speeches that a U.S. president has to give. One of these is the State of the Union Address, an annual speech that the President gives to a joint session of the U.S. Congress regarding his political concerns. In the United States Constitution, the Address is described as follows:

(1) "He [The President] shall from time to time give to the Congress Information of the State of the Union, and recommend to their Consideration such Measures as he shall judge necessary and expedient." (Article II, Section 3 of the United States Constitution)

While this address has been given ever since George Washington was inaugurated, there are some characteristics to be noted. First, although most of the addresses are

given in the form of a speech, those in the 19th to the early 20th centuries were conveyed as written reports. Secondly, the U.S. Constitution directs the president to convey his message to Congress, but the spread and development of broadcasting technologies like radio and TV, seem to have changed the targeted audience. That is, before the development of these technologies, the audience was just members of congress. However, since Warren Harding's (1921−1925) address in 1923, which was the first State of the Union Address that was broadcast to the public via the radio (Kaid, 2007: 696), the targeted addressees have included not only political members, but also millions of Americans. Last, although the addresses are expected to be given once a year, in the case of the death of a president or if a president had to resign before his report of the year, the new president would give his own address. For example, William Henry Harrison (1841−1841) passed away a month after his inauguration, hence he did not give a State of the Union Address. Instead John Tyler (1841−1845), who took over the presidency, gave his own speech.

As described above, there are some difficulties in analyzing the State of the Union Addresses. Nonetheless, public speeches adhere to certain topics or themes to convey some messages to an audience. Such topics may not be made explicit, but native speakers can understand what the addresser is speaking about. This is because a "topic is the most important situational factor influencing vocabulary choice; the words used in a text are to a large extent determined by the topic of the text" (Biber and Conrad, 2009: 46). Therefore, scrutinizing how words are used in the State of the Union Addresses may reveal the main topics of the speeches or the main political concerns of the presidents. Following this assumption, I have two goals for this study: (i) what were/are the main political interests of the U.S. presidents? That is, how have the legislative concerns changed over two centuries?; (ii) what causes the changes of interest? I will argue that, with one of the topic modeling techniques called the latent Dirichlet allocation (LDA) proposed by Blei et al. (2003), the presidents' political concerns are divided into four topics, and the turning point is strongly influenced by world-wide wars.

2. Literature Review

The U.S. State of the Union Addresses have attracted a lot of attention from

researchers (Bonnefille, 2008; 2013; Crockett and Lee, 2012; Herz and Bellaachia, 2014; Tung, 2014; *interalia*). Previous literature aimed at investigating addresses from various areas of expertise, and many of them used quantitative techniques. Although there is no previous research that I have seen that investigated my specific questions, some additional studies are worth considering. Thus, this section summarizes two important previous studies and emphasizes the importance of having a diachronic perspective.

2.1 A text-mining approach to the Addresses

First, and most relevant to us, a topic analysis was done by Crockett and Lee (2012). They examined speeches from 1989 to 2011. In these 23 addresses, they found seven topics associated with "war and terror" (which included words such as *Hussein*, *Saddam*, and *Soviet*); "economics and finance" (which included terms such as *bank*, *trillion*, and *small-business*); and "inspiration of the nation and its people" (which included *hopeful*, *homeland*, and *pension*). Note that they provided no further labels nor a result of their experiments for the rest of the topics obtained.

Tung (2014) conducted another text-mining study. His research aim was two-fold: to identify the overall lexical trends in the State of the Union Addresses, and to find differences in word usage between the two major political parties, namely Republicans and Democrats. For the first research question, he mentioned that because the data collection ranges over 200 years, discovering word usage patterns in the addresses was difficult. However, when focusing on the addresses given from 1961 to 2014, some tendencies were observed. That is, Republicans used words related to external issues, such as *war, terrorists, Iraq*, and *Saddam*, inferring that these emphasize international relations. On the other hand, Democrats tended to use terms related to economics and finance (e.g., *companies, businesses, payments*, and *employment*) and related to social affairs (e.g., *education, students, wage*, and *jobs*). In other words, three major topics that the presidents tended to speak about are war and terror, economics and finance, and social affairs.

2.2 Diachronic perspective

Though the studies reviewed in the last subsection are interesting and thought-provoking, I shall point out the necessity of a diachronic transition in topics, which has

been neglected in the previous literature.

In Crockett and Lee (2012), the seven topics they found seem to be certain time- or period-specific issues. For example, regarding the topic of war and terror, proper nouns like *Soviet* were certainly related to the Cold War, whereas the term *Saddam Hussein* was related to Hussein's trial for genocide, which made the headlines in the mid-2000s. In other words, while it is true that both words are related to the topic of war and terror, Saddam Hussein's execution was a decade and a half after the Cold War ended. The same can be said of Tung's analysis. His results showed that words frequently used by Republicans included *Iraq, Iraqi, terrorist(s), Saddam*, and *Hussein*, which all pertained to September 11, 2001 and the Iraq War. Words used by Democrats involved *Vietnam* and *Soviet*, referring to the Cold War. As Tung himself recognized, 9/11 and the execution of Saddam Hussein took place when G. W. Bush (2001–2009) was the president. Similarly, the worst period of the Cold War, the Cuban Missile Crisis in 1962, and the first half of the Vietnam War, were overseen by Democrat presidents. Hence, the wars and terror happened at completely different times.

Considering the previous literature and the diachronic approach to the addresses, studies mentioned above emphasize the importance of investigating the changes in word and phrase usage. Put differently, in order to investigate our research questions introduced above—namely, how the presidents' political interests have changed over time and what motivates them to have such ideas—I have to analyze topics that the presidents have expressed concern about, and the topics revealed must be examined in chronological order.

With this assumption, one of the topic modeling methods, the LDA proposed by Blei et al. (2003), was applied to the collection of the State of Union Addresses.

3. Methods

This section will introduce the corpus and method this study used. Note that all the statistics and figures are proceeded by R (3.4.3) except for concordance lines and simple frequency counts.

3.1 Corpus

In this study, I collected as many speeches as possible from various web

Table 1: The information of the UA Corpus

Number of files	Tokens	Types	STTR (per 10,000)
229	11,915,915	156,860	27.28

resources. Because the manuscripts were obtained from different web sites, many symbol and spelling variations were found. Thus, non-computer friendly symbols are replaced with the HTML format and the spelling variations are standardized into the current American English orthography. For example, some manuscripts mistakenly used hyphens instead of en dashes, so I changed them into en dashes, represented as <–>. Other symbols, such as quotation marks and ampersands, were also replaced with the HTML format, which was/is enclosed in angled brackets. Next, I obtained a corpus as described in Table 1, which was calculated by CasualConc (Imao, 2015). This corpus will be called the United States presidents' State of the Union Address corpus (abbreviated as the UA Corpus).[1]

3.2 Latent Dirichlet allocation

As already explained, the aim of the current study is to investigate how the topics in the UA Corpus have changed. To achieve this goal, I used the LDA proposed by Blei et al. (2003) because its basic premise perfectly coincides with our research question. As Blei et al. (2003: 996) stated, "[t]he basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words." This idea is quite compatible with corpus linguistics research, as I quoted Biber's and Conrad's explanation.

Another important hypothesis under this method is that the LDA is based on the "bag-of-words" assumption—word orders or document orders are ignored, and all words within a document are exchangeable. This means that even though the UA Corpus is compiled with diachronic data and the current research question is to investigate transitions in topics over time, the method itself has nothing to do with a diachronic perspective.

How, then, are the topics obtained? First, stop-words should be removed. Topics are what addressers are speaking about, and thus topics are basically composed of content words, not functional counterparts. Hence, I slightly modified a stop word list used in Tabata's (2017) study and removed the words. Then, the number of topics to be

computed, which is at researcher's own discretion, must be decided (for a generous introduction to the LDA, see Kuroda (2017), Tabata (2017) and Schöch (2016)). Thus, in this study, the number of topics were calculated by changing the number ranging from ten to seventy at every five intervals. Then, the sets of words contributing to each topic were scrutinized, and the set with 35 topics seemed to capture the distribution of topics.

Note that because the aim of this study is to uncover the diachronic change in political interests of the presidents, topics that were observed only in specific periods were discarded. For example, Topic 33 was observed during William McKinley's (1897–1901) and Theodore Roosevelt's (1901–1909) administrations, but no other presidents talked about this topic. It is doubtful that this topic represents a transition in political interests in the long term. Thus, such topics were excluded from this study.

The last thing to be mentioned in relation to the LDA with the current study is that the LDA calculates the degree to which the speech belongs to the topics. In other words, all the files belong to every topic, but they differ in how high their contributions were to the topics. It would be best if all the probability ratios could be used; however, because the current research interest was to uncover the political interests of the presidents, the topics that were considered as the most likely topics of the documents were used. What I mean by "the most likely topics" are topics whose probability rates as calculated in the LDA exceed 0.3 in various or stable periods. In this study, topics that lasted over 15 years were considered as the most likely topics.

Some readers may think that 0.3 in the probability rate is too low. However, the average of the overall probability is 0.002, and topics that exceed the threshold of the probability rate, appearing in various or consecutive periods, are quite limited. Thus, I think that my standard is appropriate, albeit slightly arbitrary.

3.3 Concordance lines and n-grams

After labeling each topic obtained from the LDA, this paper goes into further detail regarding what motivated the presidents to be highly concerned about the topics. This question will be proceeded by scrutinizing how words that highly contribute to the topics are used in addresses by focusing on quad-grams recurrently used with respect to raw-frequency. The reason why I used quad-gram, and not bi-, tetra-, or hexagrams is two-fold: semantic contents and frequencies.

First, the longer the collocations are, the more meaningful they are, and thus it is very easy to guess in what situation or context the presidents used them. On the contrary, short collocations are rarely meaningful enough to grasp the way the clusters are used. To be more specific, compare the hexagrams and bigrams of *American* attested in the UA Corpus, which are shown in Table 2. The table shows that the bigrams are too short and that hexagrams are very precise in their contexts. In this sense, longer collocations enjoy their reputation of being meaningful. On the other hand, the frequency information reverses their situations, in that the longer the collocations are, the less frequent they are. The largest number of the hexagrams in Table 2 is, at most, 6. Then, is it significant enough to conclude that the expression *of the claim of American citizen* is more important or frequent than other collocations? I would say that it is not, or at most, that it is quite a difficult question to answer. On the contrary, the bigrams show that *the American* is by far the most frequently found expression, and it should be safe to say that it is one of the most frequent collocations on the list. Thus, it is much easier to distinguish key phrases in shorter collocations than in longer ones.

These two mutually exclusive issues are resolved (but not without any issues) by assessing quadgrams. Thus, this study will focus on quadgrams, and show their concordance lines to explore the research question. Note that all the frequency counts on collocations and concordance lines presented in this and later sections are generated by CasualConc.

Table 2: Comparing frequencies of hexagrams and bigrams of *American*

No.	Hexagram	Freq.	Bigram	Freq.
1	of the claim of *american* citizen	6	the *american*	872
2	the live and property of *american*	5	of *american*	420
3	live and property of *american* citizen	4	*american* people	390
4	the bureau of the *american* republic	4	*american* citizen	195
5	the central and south *american* state	4	every *american*	122
6	the character of the *american* people	4	a(n) *american*	101
7	the claim of *american* citizen against	4	to *american*	72
8	the spirit of the *american* people	4	*american* republic	64
9	a decent home for every *american*	3	*american* family	50
10	*american* citizen against the government of	3	*american* vessel	49

4. Results and discussion

Based on the processes explained in the previous section, I obtained four topics: 6, 10, 27, and 34.[2] A smoothed diachronic figure is given in Figure 1, which illustrates that Topics 6 and 34 were the major concerns of the presidents at the beginning of the United States. However, as their importance decreases, Topics 10 and 27 gain their prominence, and the transition of Topic 27 is more radical than that of Topic 10.[3] Then, our question is, what does each topic represent? This is well-represented through a word cloud, using words whose word weight is 0.0008 or higher, and such words are plotted in Figure 2 to 5,[4] representing each topic; the following sections will deal with the topics in turn.

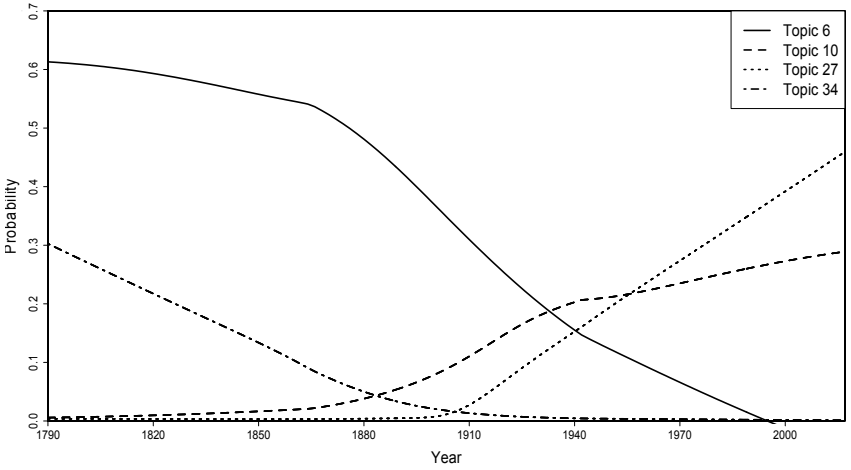


Figure 1: Diachronic transition of the topics

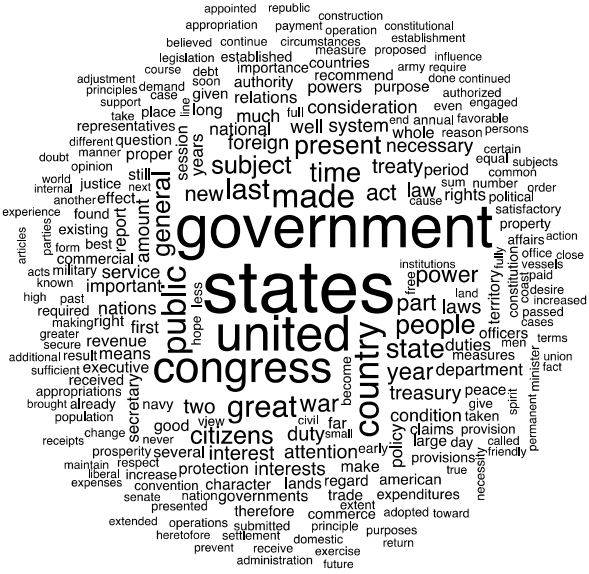


Figure 2: Word cloud on Topic 6

Figure 3: Word cloud on Topic 34

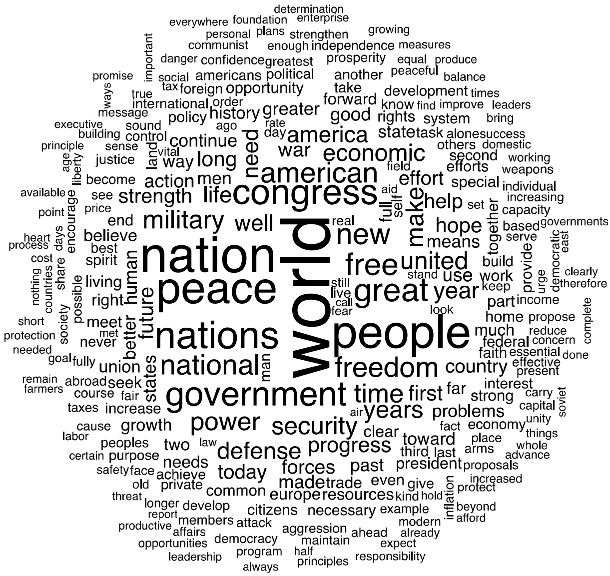


Figure 4: Word cloud on Topic 10

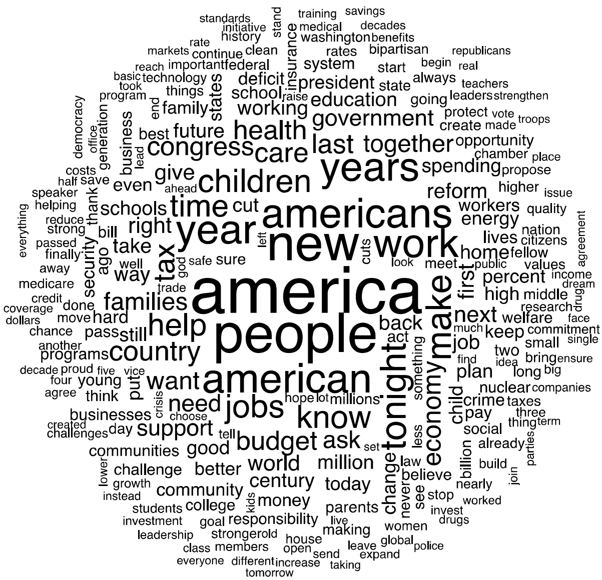


Figure 5: Word cloud on Topic 27

4.1 Topic 6 (1790−1920s)

This topic was considered to be a quite important topic at the beginning, as no other topics received more than 0.5 in Figure 1. Figure 2 shows that words related to domestic affairs, such as *congress, right(s), law(s), constitution(al), territory, duty, duties, government(s), country,* and *parties*, are frequently observed in this topic. The appearance of these words indicate that the topic is related to politics. Then, the question arises: whose politics are they, the United States' or the international community's?

In order to understand the topic more precisely, I looked at how these words were used.[5] One of the most frequently used words in the list above was *right(s)*, which appeared nearly 1,500 times out of 2,217 total frequencies in this period, and its most frequent quadgram *the right of the*, was found mostly in this period (64 times out of 69 times in total). As the concordance line in Figure 6 shows, words that follow the collocation, namely the possessor of the right, are either political organizations such as *(united) states, government*, and *party*, or those related to citizens such as *people, Indian* and *minority*. These collocational patterns indicate that the rights referred to in this period were related to domestic affairs. Expressions related to U.S. internal situations in this topic were found not only in the collocation of *right*. Let us consider two more salient examples, *the constitution of the*, the second-most commonly used quadgram of *constitution*, and *the duty of the* or *the duties of the*, the most frequent quadgram of the lemmatized form of *duty*, in turn. Figure 7 shows the linear sequence of *the constitution of the*, and it illustrates that the noun phrases following the quadgram were dominantly occupied with either *united states* or *states*, which were mostly found in the period between the 1700s and 1920 (26 times out of 28). Figure 8 also exemplifies presidents' interests, in that the presidents paid special attention to what the U.S. government or the president must work on, because the *of* phrase mostly

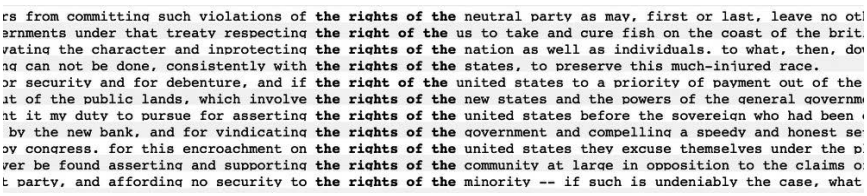


Figure 6: The concordance lines of *the rights of the*

important state of north carolina to **the constitution of the** united states (of which offi
ise the legislative power granted by **the constitution of the** united states "to lay and co
ity to preserve, protect, and defend **the constitution of the** united states", on you, gent
us depository of american happiness, **the constitution of the** united states. let them cher
the constitution of the judiciary, experimental and
mblies deriving their authority from **the constitution of the** state. each is sovereign wit
he inhabitants of the united states. **the constitution of the** united states requires that
ny would be vain and ridiculous. but **the constitution of the** united states imposes on the
by **the constitution of the** united states it is provided
ns, to which they are entitled under **the constitution of the** united states. they should h
e conformable to the requisitions of **the constitution of the** united states, i recommend t
the constitution of the united states provides that

Figure 7: The concordance lines of *the constitution of the*

rency it has been assumed that it was **the duty of the** executive not only to suppress insur
is, should be persisted in, it will be **the duty of the** united states to resist its executio
ey have taught. in the meantime it is **the duty of the** government, by all proper means with
which has so long existed and render **the duty of the** president plain in executing its pro
ect a member of congress, it shall be **the duty of the** president to cause a census of the i
is arises from an obstacle which it is **the duty of the** spanish government to remove. whilst
guidance and protection. whilst it is **the duty of the** president "from time to time to give
positions, prepare the way. i hold it **the duty of the** executive to insist upon frugality i
itution of the united states makes it **the duty of the** president to recommend to the consid

Figure 8: The concordance lines of *the duty|duties of the*

included *president, government,* or *united states*. Note that the phrase *the united states* was used in one of two ways, either referring to the country (namely, the United States) or to two or more states being in association with each other (hence, united states).

All of these collocations lead us to conclude that Topic 6 can be labeled as the internal issues of the federal government. It is relatively straightforward as to why this topic was the main concern of the presidents; the country was founded in 1776, and the government had to develop a constitution and laws making the President's role explicit to the U.S. Congress. Therefore, the presidents frequently referred to the topics related to the federal government.

4.2 Topic 34 (1790s−1870s)

Topic 34 was salient in the first century of the United States. Figure 3 shows words that contribute to making up this topic, which show some tendencies. That is, this topic includes words pertained to regions, nationalities, and races such as *Britain, Spain, French, Florida, tribes* and *Indians*. Furthermore, we can find words related to water, such as *lake, ocean, waters*, and *navigation*. Then, does this topic consist of two different topics?

Let us take a look at a few of the words listed above. The collocational patterns of

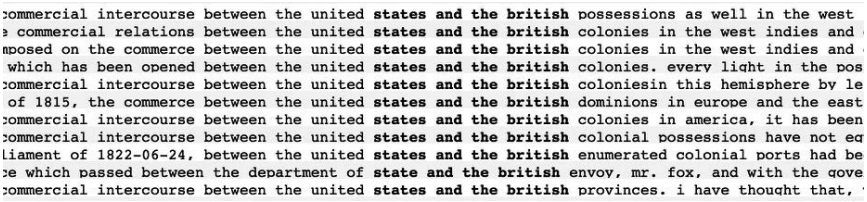


Figure 9: The concordance lines of *states and the British*

Figure 10: The concordance lines of *between Spain and*

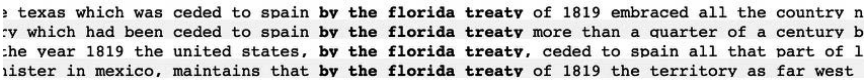


Figure 11: The concordance lines of *by the Florida treaty*

the country names reflect U.S. history. Figure 9 shows one of the frequently used collocations of *British*, and as it reads, expressions such as *colonies*, *possession*, and other related phrases follow after *United States* and *British*. Similarly, the collocations of *Spain*, demonstrated in Figure 10, indicate that there was a territorial dispute between Spain and the United States. These concordance lines suggest that there used to be many Spanish and British territories on the North American continent at that time and that the presidents were highly concerned about relationships with the two countries. This is also confirmed by the quadgram of *Florida*, which is given in Figure 11.

Another perspective is found by looking at words related to water. Consider the most frequently observed quadgram of *lake*, as given in Figure 12. As the concordance lines show, the fact that areas between the Lake of the Woods and the Rocky Mountains used to be a British territory was one of the main concerns of the presidents. In other words, there were lots of territorial disputes around the United States, and the presidents paid special attention to the issues.

As the concordance lines in Figure 12 show, *lake* is strongly tied to the territorial

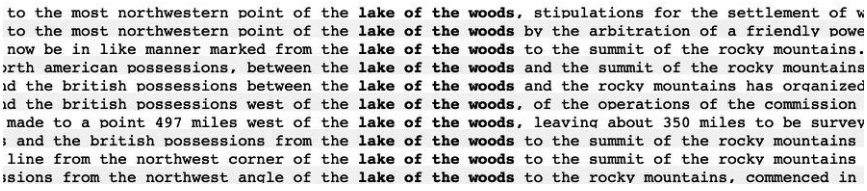


Figure 12: The concordance lines of *Lake of the Woods*

disputes with Spain and Britain. It is interesting to note that words related to water are often used to refer to a relation between the United States and other countries. For instance, many examples of *to the Pacific Ocean*, the most frequently observed quadgram of *ocean*, occur by referring to either country or racial names, as in (2) to (4):

(2) James Monroe's speech in 1818 (Democratic-Republican Party)
[I]t has been necessary during the present year to maintain, a strong naval force in the Mediterranean and in the Gulf of Mexico, and to send some public ships along the southern coast and *to the Pacific Ocean*. By these means amicable relations with the Barbary powers have been preserved, our commerce has been protected, and our rights respected.

(3) Andrew Jackson's speech in 1846 (Democratic Party)
[F]rom central America I have received assurances of the most friendly kind and a gratifying application for our good offices to remove a supposed indisposition toward that government in a neighboring state. [...] Our treaty with this republic continues to be faithfully observed, and promises a great and beneficial commerce between the two countries - a commerce of the greatest importance if the magnificent project of a ship canal through the dominions of that state from the Atlantic *to the Pacific Ocean*, now in serious contemplation, shall be executed.

(4) James Knox Polk's speech in 1846 (Democratic Party)
[O]ur laws regulating trade and intercourse with the Indian tribes east of the Rocky Mountains should be extended *to the Pacific Ocean*.

Furthermore, Figure 13 shows the most frequently observed quadgrams *navigation, of commerce and navigation*, and, as the data show, the collocations are followed by country names many times. These two observations strongly suggest that the presidents in this period were concerned about relationships with foreign countries tied with area issues.

Having scrutinized many examples of frequently observed lexical categories, and

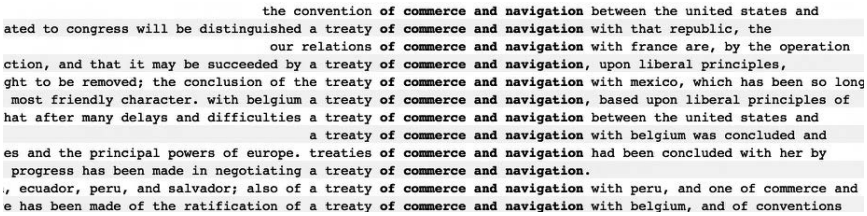


Figure 13: The concordance lines of *commerce and navigation*

having argued that the United States had territorial disputes in its early history and that presidents frequently referred to relations with others, I would like to label Topic 34 as international affairs, strongly tied with territorial disputes. Note that "international" here may be misleading, because some conflicts at that time were not strictly those between countries but were confrontations with first nations living outside of the U.S. boundary. Nonetheless, I use the term here for convenience.

4.3 Topic 10 (1880s−2010s)

The next topic, Topic 10, gradually increased its importance around the late 19th century and gained the highest score around the 1940s. Figure 4 illustrates that many military-related words, such as *peace*, *defense*, *military*, *power*, *war*, and *forces* contributed to this topic. Furthermore, the word cloud includes items referring to international relations, such as *world*, *Europe*, *Soviet*, and *countries*. These examples are enough to label this topic worldwide warfare.

In order to understand why this topic gained presidents' main political focus from the 1880s to the 2010s, let us observe some words other than the military-related words found in Figure 4. First, let us consider *communist*. The word was not very frequently used (130 times in total), because its first appearance was in Truman's speech in 1950, and not many specific collocation patterns were found. Thus, we shall take a look at words that occur within five words—the left and right windows—of *communist*. Table 3 shows the most frequently observed co-occurring words orbiting around *communist*. Figure 14 gives examples of how *communist* and *world* were used in actual manuscripts, which was mostly to evoke a negative impression or to show aggression against communism by using words such as *aggression*, *threaten*, *conspiracy*, and *painful phase*. This was apparently embodied in Ronald Reagan's expression referring

Table 3: Co-occurring words of *communist*

No.	Word	Freq.	No.	Word	Freq.	No.	Word	Freq.	No.	Word	Freq.
1	the	124	6	and	28	11	chinese	12	16	with	11
2	of	72	7	have	28	12	nation	12	17	against	10
3	be	45	8	a	24	13	world	12	18	china	10
4	to	42	9	that	17	14	aggression	11	19	threat	9
5	in	31	10	by	12	15	this	11	20	all	8

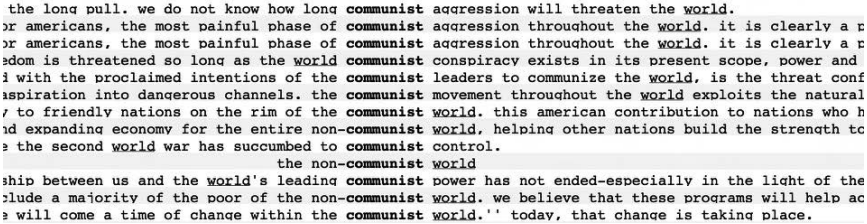


Figure 14: The concordance lines of *communist*

to the Soviet Union as the *evil empire*. Furthermore, *communist* also occurs with *Chinese* and *China* as shown in Figure 15. Given the data in Table 3 and Figure 14 and Figure 15, this topic was strongly influenced by the Cold War.

Another word that indicates that this topic was driven by the Cold War was *nuclear*. It is true that *nuclear* may refer to a nuclear plant, but most of the examples of *nuclear* occurred with military-related words such as *weapon*, *threat*, *war*, and *force*, as exemplified in Figure 16. Historically speaking, the United States competed with the Soviet Union to increase their number and the power of nuclear weapons, and in the early 1960s, the Cold War faced an urgent situation due to the Cuban Missile Crisis. Thus, it is no wonder that the presidents in this period focused on nuclear weapons.

Therefore, as the observations in the word cloud in Figure 4 and the concordance lines given in this subsection show, this international warfare topic gained the prominent focus of presidents because of the Cold War.

venth fleet no longer be employed to shield **communist** china. this order implies no aggressive intent
as a base of operations against the chinese **communist** mainland.
was required to serve as a defensive arm of **communist** china. regardless of the situation in 1950,
venth fleet no longer be employed to shield **communist** china. this order implies no aggressive intent
as a base of operations against the chinese **communist** mainland.
al methods and backward course of events in **communist** china. in these continuing efforts, the free
neral assembly, its secretary-general is in **communist** china on a mission of deepest concern to all
. in the release of our fifteen fliers from **communist** china, an essential prelude was the world
ry has continued to withhold recognition of **communist** china and to oppose vigorously the admission
ll-out bombardment of quemoy restrained the **communist** chinese from attempting to invade the off-
ill our relations with the soviet union and **communist** china. we must never be lulled into believing

Figure 15: The collocational patterns of *communist and China | Chinese*

and mobility of our present conventional and **nuclear** forces and weapons systems in the light of present
manning, and directing a truly multilateral **nuclear** force within an increasingly intimate nato allianc
ting new nations to master the black arts of **nuclear** war -- and if they are willing to turn their energ
sition of having to answer every threat with **nuclear** weapons or nothing.
diture of more than $15 billion this year on **nuclear** weapons systems alone, a sum which is about equal
reaty -- to demonstrate both the futility of **nuclear** war and the possibilities of lasting peace.
nal to launch a nuclear attack or to use its **nuclear** power as a credible threat against us or against o
is at all possible, in both conventional and **nuclear** weapons and defenses. i thought we were making som
ch an agreement that will halt the spread of **nuclear** weapons. on the basis of communications from
ast, the conflict in vietnam, the dangers of **nuclear** war, the great difficulties of dealing with the
we will maintain a **nuclear** deterrent adequate to meet any threat to the secur
rve our interests and minimize the threat of **nuclear** confrontation.
in an era where the strategic **nuclear** forces are in rough equilibrium, the risks of conf

Figure 16: The concordance lines of *nuclear*

4.4 Topic 27 (1950s–2010s)

Lastly, consider Topic 27. The word cloud in Figure 5 shows that this topic includes words that can be clustered in various ways, such as family-related issues (*child(ren), education, schools, college, family*), labor (*businesses, jobs, income,* and *companies*), and health care (*medical*, *insurance, health*, and *Medicare*). With these observations, I would like to label this topic social welfare, and consequently, more lexical items are found to be related to this topic, such as *tax(es), woman, women, social, budget, vote, safe, energy*, and *right*.

Let us take a look at some words that seem to be unrelated to social welfare. The word *security* is found in the word cloud, and this may evoke national security related to warfare. It is true that *security* is often used to refer to national defense, as its most frequently observed quadgrams are *the security of the, to the security of*, and *for the security of*, but a collocation that is strongly tied to this topic, namely *the social security system*, is found to be the fifth-most frequent quadgram. Furthermore, if we restrict ourselves to the bigram and do not use the quadgram, *social security* occurs only in this period (199 times).

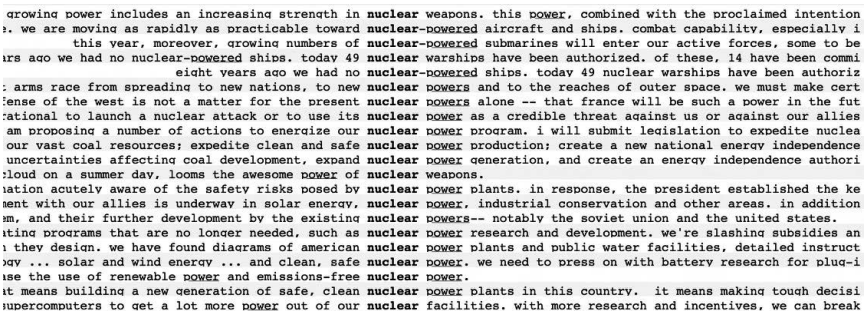


Figure 17: The concordance lines of *nuclear*

Another controversial word is *nuclear*, in that it is also a keyword in the last topic, namely worldwide warfare. One of the frequently used bigrams, *nuclear power*, presents an interesting change in the context of the bigram (see Figure 17). The first half (from 1955 to 1977) of the collocations shows that the presidents used *nuclear power* to refer to a source of weapons, whereas the later presidents (1981 to 2011) tended to use *nuclear power* to refer to electricity. This change reflects differences in the two topics described above and here, namely, worldwide warfare and social welfare.

Why did social welfare gain the primary attention of presidents? First, the mid-1960s was a period of civil rights campaigns in the United States, as various facts show (e.g., in 1963, Martin Luther King Jr. gave his famous “I have a dream” speech; Michael Marrington published *The Other America*, which discussed the existence of economically handicapped people such as the elderly and minorities; and Medicaid was legislated). Furthermore, the growing feminism campaign motivated a change in the way women were treated. That is, inequality in wages between males and females was legally eliminated in 1963. This is probably why *woman* or *women* is made explicit in the context of serving the country in one of the most frequent quadgrams, *men and women who*, as exemplified in Figure 18. Thus, it is no wonder that the topic of social welfare starts to gain presidents’ attention at this time. Second, environmental issues became known to the world. As various leading industrial areas are concentrated in the United States, the consumption of energy that causes destruction of the environment must be dealt with. Hence, the topic also covers energy affairs during this time.

Summarizing this section, I have shown that the presidents’ main political

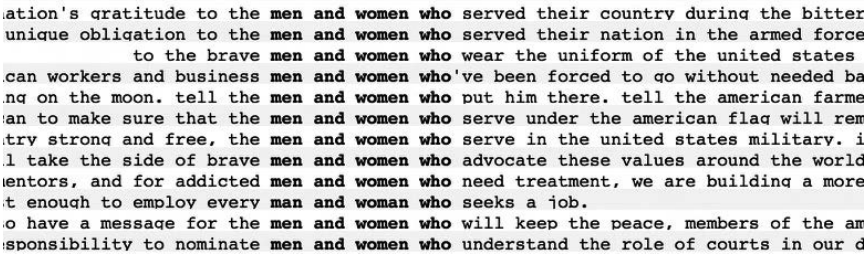


Figure 18: The concordance lines of *men and women who* and *man and woman who*

concerns have changed among four topics. The Presidents of the first century who were involved in foundation of a nation, were mainly concerned about internal or domestic affairs. At the same time, the presidents of the early days of the United States also focused on international relationships, especially emphasizing territorial issues. After World War II, the political concerns of presidents gradually changed, and worldwide warfare and social welfare became the presidents' primary considerations.

Note that the presidents were concerned about more than one topic in the same period. As Figure 1 shows, the moves in Topics 6 and 34, and those in Topics 10 and 27, are quite similar, in that the time the first two topics falls increases the other two topics. Put differently, Topics 6 and 34 are negatively correlated to Topics 10 and 27. There is no clear-cut boundary of exactly when the relations changed, but the mid-1930s, namely the time when the United States attained the world's supremacy, seems to be the threshold during which Topics 10 and 27 took the primary considerations of the presidents.

5. Conclusion

In this study, I have demonstrated how the presidents' primary concerns have changed over two centuries by applying the LDA to the State of the Union Addresses and showed that four main topics are obtained: internal issues related to the federal government, international affairs tied with territorial disputes, worldwide warfare, and social welfare. Lastly, I proposed that the transition in the topics occurred around the mid-1930s. In other words, the presidents' main political concerns were influenced by whether or not the United States had attained the status of a world leader.

There are a few remaining issues to be addressed. First, as was stated earlier, not all of the speeches were given as oral presentations; some were merely submitted written reports. Assuming that register variants do not affect the topics, I intentionally ignored the difference in register. Nonetheless, it may have some influence on our conclusion, and thus, register variations should be given further concern. Secondly, there are some criteria that were arbitrarily made, such as selecting the number of topics and which topics were to be scrutinized. Thus, it is necessary to find less ad-hoc methods (a heat map may help us resolve this issue). Lastly, since 1923, the targeted audience has changed, and this may also have affected the transitions of topics. More specifically, the social welfare topics may be associated with the change in target audience, though I do not have enough confidence to add this explanation for Topic 27.

Acknowledgment

An earlier version of this paper was presented at JAECS 43, which was held at Kwansei Gakuin University. I would like to thank the insightful comments from the audience. I am also grateful to Tomoji Tabata for providing me with his stop word list and to Roger Prior for his invaluable comments. Lastly, I would like to thank Yoshiyuki Nakao, the editor of the English Corpus Studies 25, and three anonymous reviewers for their insightful comments. Of course, any remaining errors are my own. This study was supported by a research grant from University of Kitakyushu.

References

Biber, D. and S. Conrad (2009) *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Blei, D., A. Ng, and M. Jordan (2003) "Latent Dirichlet Allocation." *Journal of Machine Learning Research 3*: 993–1022.

Bonnefille, S. (2008) "When Green Rhetoric and Cognitive Linguistics Meet: President G. W. Bush's Environmental Discourse in his State of the Union Addresses." *Metaphorik.de 15*: 27–61.

Bonnefille, S. (2013) "Energy Independence: President Obama's Rhetoric of a Success Story." *Research in Language 11*: 189–212.

Crockett, S. and C. Lee (2012) "Does It Matter Qhat They Said? A Text Mining Analysis of the State of the Union Addresses of USA Presidents." *13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*: 77–82.

Herz, J. and A. Bellaachia (2014) "The Authorship of Audacity: Data Mining and Stylometric Analysis of Barack Obama Speeches." In Stahlbock, R., G. M. Weiss, M. Abou-nasr and H. R. Arabnia (eds.), *Data Mining: Proceedings of the 2014 International Conference on Data Mining (Worldcomp Iinternational Conference Proceedings 2014)* Nevada: CSREA Press, pp. 148–154.

Imao, Y. (2015) CasualConc (Version 2.0.5) [Computer Software] URL: https://sites.google.com/site/casualconc/

Kaid, L. L. (2007) "Radio, Politics and." In Kaid, L. L. and C. Holtz-Bacha (eds.), *Encyclopedia of Political Communication.* California: SAGE Publications, pp. 696–697.

Kuroda, A. (2017) "Quantitative Analysis of Literary Works: Novels of Sir Arthur Conan Doyle." In the Institute of Statistical Mathematics (ed.). *Text-mining and Digital Humanities*. Tokyo: the Institute of Statistical Mathematics, pp. 55–70.

Schöch, C. (2016) "Topic Modeling Genre: An Explanation of French Classical and Enlightenment Drama." *Digital Humanities Quarterly 11*, URL http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html

Tabata, T. (2017) "The Semantic Structure of the FLOB Corpus: Using Probabilistic Models for Corpus Description." In the Institute of Statistical Mathematics (ed.). *Text-mining and Digital Humanities*. Tokyo: the Institute of Statistical Mathematics, pp. 1–18.

Tung, J. (2014) Text Mining Analysis of State of the Union Addresses: With a Focus on Republicans and Democrats Between 1961 and 2014 URL: https://statoftheart.files.wordpress.com/2014/05/text-mining-analysis-of-state-of-the-union-addresses.pdf

Notes

1. The dataset is available at <https://drive.google.com/file/d/0B27VXLzIM-qhNmxNVnNsTk84bEk/view>
2. The topic numbers are computed through the LDA, their orders have nothing to do with diachronic counterparts.
3. The movement of the topics indicates that the U.S. presidents do not have any everlasting political philosophy that is consistent through 240 years.
4. Readers may think that 0.0008 as a cutting point looks too low. However, the average of the overall word weight is 0.00004, and the maximal value is 0.04. I did trials-and-errors many times to find a value to plot visible word clouds, and 0.0008 was the lowest relatively visible value. Thus, I would like to say that the value is appropriate.
5. It should be natural to examine words represented in the center of Figure 2, such as *united*, *states*, and *government*, however the quadgram of such words show that they are mostly used as *government of united states of America*, and no insightful quadgrams were found. Thus, being certainly quite important to this topic, such words do not provide any answers to the current question, and will not be further investigated.

（木山　直毅　北九州市立大学　Email: n-kiyama@kitakyu-u.ac.jp）

「研究ノート」

Construction of Medical Research Article Corpora with AntCorGen: Pedagogical Implications

Motoko ASANO

Abstract

This paper examines the usefulness of a novel corpus generating tool AntCorGen (Anthony, 2017a) for the compilation of medical research article corpora for use in pedagogical settings. A corpus comprising approximately 1,500 *PLOS ONE* medical research article abstracts was compiled successfully using AntCorGen. A larger corpus was also built using 400 research articles encompassing approximately 1.73 million words from four disciplines of medicine and was explored from the viewpoint of English for Specific Purposes (ESP). The findings suggest much promise for the use of corpora in the classroom.

1. Introduction

With the emergence of "English as the *lingua franca* of scientific communication in general[,] and of medicine in particular" (Salager-Meyer, 2014: p. 49), English is being used in approximately eight million health-related peer-reviewed articles published globally each year (Salager-Meyer, 2014). Medical students and researchers need to be able to efficiently use English to become members of their disciplinary discourse communities (Mauranen, 2017).

Linguistic support for apprentice scientists has been provided along with research on medical discourse (Maher, 1986). According to Salager-Meyer (2014), genres within ESP contexts that have been studied most extensively are research articles (RAs) and case reports. Following the framework of rhetorical movement or "moves" in RA introductions, which was termed "*Create a Research Space* (CARS) model" (Swales, 1990, p. 140), Nwogu (1997) analyzed 15 medical research articles from international journals such as *The Lancet* and suggested 11 moves. Salager-Meyer (1989) analyzed

linguistic features of 51 medical texts including RAs, case reports, and editorials. In all these studies, small, specialized corpora were built and used. This approach has been considered legitimate as Lee (2008, p. 94) maintains that "discourse analysts who work with specialized discourses can benefit from compiling their own corpora and applying some of the techniques of corpus-based linguists to support their analyses."

For teaching, corpora have been "extremely useful for ESP teachers in that they are able to show how language is used in the context of particular academic genres" (Paltridge, 2013, p. 351). The findings from textual analyses have been incorporated into various textbooks for RA writing (Feak & Swales, 2012; Nakatani & Bucsis (Ed.), 2016; Noguchi, Matsuura, & Haruta, 2015).

Recently, many science and engineering RAs come from China and Japan (Flowerdew & Wang, 2017). For science students writing in their second language, "a certain extent of language re-use from other texts is *acceptable*" (Flowerdew & Li, 2007, p. 442). Flowerdew and Li (2007) maintain that "re-use of formulaic structures at the syntactic level and formulaic chunks at the lexical level are basic learning strategies upheld by corpus-based pedagogy, while the formulaicity of various texts at the rhetorical level has been the foundation of genre-based pedagogy" (p. 460).

In classrooms, learners use corpora to discover linguistic patterns. Such "inductive approaches to the learning of grammar and vocabulary" (Jones & Dimant, 2013, p. 395) are known as data-driven learning (DDL) (Johns, 1991). The DDL approach has been used for students from disciplines such as educational technology (Lee & Swales, 2006) and biological sciences (Noguchi, 2004); learners download articles from their target journals and create mini corpora. Concordance texts are observed for "the simulation of inductive learning strategies" (Johns, 1991, p. 30). However, building corpora can be an arduous task even for those well versed in English for research publication purposes (Swales, 1990; Nwogu, 1997; Maswana, Kanamaru, & Tajino, 2015).

To offer help with corpus building, Anthony (2017b, p. 71) developed the novel freeware AntCorGen (Anthony, 2017a), which enables rapid, automatic generation of discipline-specific corpora from articles in the *PLOS ONE* research database, thus "making the tool ideally suited for use in Data-Driven Learning." This paper reports on the use of AntCorGen to build discipline-specific corpora of *PLOS ONE* medical RAs and examines possible applications for classroom use.

2. Examination—The CAPHYRA corpus

2.1 Corpus construction

A corpus of cardiovascular physiology RA abstracts (CAPHYRA) was prepared using AntCorGen (Version 1.1.0), which was the latest version at the time of the study. The field of cardiovascular physiology was chosen because physiology is one of the basic fields of medicine (Caze, 2011) and because cardiovascular RAs have been used for linguistic studies (Coates, Sturgeon, Bohannan, & Pasini, 2002). As AntCorGen indicated that the number of physiology RAs was 50,101, the RA category was limited to "cardiovascular physiology," and only the abstract portion of the RAs was collected using the following procedure:

1. A new folder was created on the desktop of a computer.
2. AntCorGen was launched by left clicking the AntCorGen icon and selecting run as an administrator, which was a safer way to avoid freezing of the software during the operation.
3. The corpus storage folder, which was prepared in procedure 1, was selected to store the files.
4. The research article category was chosen by checking the [cardiovascular physiology] box.
5. The target field of collection was selected by clicking the abstract box.
6. The [create corpus] button was pressed to download the article abstract texts.

The CAPHYRA corpus, comprised of 1,551 abstracts, was built within a few seconds. The file name of each abstract text, such as "_10_1371_journal_pone_0012983.txt," represents the URL of the original RAs. Therefore, DDL instructors and students should be able to easily locate the original RAs.

2.2 Methods for examining the CAPHYRA corpus

The CAPHYRA corpus was examined using the concordance software AntConc (Anthony, 2014) because the tool has been used in RA writing classrooms (Noguchi, 2004) and is recommended in a textbook (Feak & Swales, 2012). The types and tokens were counted. The most frequent words were identified, and the most frequent four-

word expressions, called “4-grams” (Nesi, 2013, p. 418), were extracted because “[m]ost researchers have chosen to examine four-word combinations” and because “[t]hey usually reveal more about the genre of the corpus than its topic” (p. 418). Concordance lines were obtained for some of the most frequent words. As Hunston (2013: p. 158) indicates that concordancing tools “only find and organize the data. Interpretation is a human activity,” the concordance lines were observed to identify patterns.

2.3 Results and Discussion

The CAPHYRA corpus had 14,954 word types and 386,966 tokens. The ten most frequent words and the normalized frequency (NF) are shown in Table 1. The two most frequent words were *the* and *of*. Marco (2000, p. 67), who analyzed collocational frameworks in medical RAs, suggested that “[t]he most common frameworks in [their] corpus are: *the ...of* (e.g., *the number of*), *be ... to* (e.g., be similar to), *a ... of* (e.g., *a variety of*)....”

Table 1: The most frequent words in the CAPHYRA corpus: Normalized per 1000 words (1551 CArdiovascular PHYsiology Research Article Abstracts)

Rank	Word	NF
1	the	40.3
2	of	39.3
3	and	38.7
4	in	33.7
5	to	16.5
6	a	15.8
7	with	10.3
8	that	9.8
9	by	8.1
10	was	7.8

The ten most frequent 4-grams are shown in Table 2. Among them, *vascular endothelial growth factor*, *endothelial growth factor vegf*, *human umbilical vein endothelial*, *umbilical vein endothelial cells*, and *of vascular endothelial growth* were highly specific technical expressions related to “the cell layer that lines the blood vessels” (Pocock, Richards, & Richards, 2006, p. 57). Two of the remaining five 4-grams were *in vitro and in* and *vitro and in vivo*. According to the American Medical

Association (AMA) Manual of Style (Iverson, Christiansen, Flanagin, & Fontanaroas. 2007, p. 925), the terms *in vitro* and *in vivo* are "considered to have become part of the English language" and "[i]talics are not used" in medical articles.

The remaining three were *in this study we, this study was to,* and *of this study was.* These are typical hint expressions which are likely to have been used intentionally by the writers of the abstracts to realize their rhetorical purposes (Tojo, Hayashi, & Noguchi, 2014; Mizumoto, Hamatani, & Imao, 2017).

Table 2: The most frequent 4-grams in the CAPHYRA corpus

Rank	Four-word expression	Frequency
1	vascular endothelial growth factor	293
2	endothelial growth factor vegf	173
3	in this study we	145
4	in vitro and in	118
5	vitro and in vivo	112
6	human umbilical vein endothelial	89
7	umbilical vein endothelial cells	77
8	this study was to	76
9	of this study was	72
10	of vascular endothelial growth	72

2.4 Pedagogical Implication

The concordance lines showed that the phrase *vascular endothelial growth factor* has no article in the third, fourth, and sixth sentences (Figure 1) and is followed by the abbreviation in all sentences. The abbreviation *VEGF* and its expanded form *vascular endothelial growth factor* appear in "the list of clinical, technical, and other common terms" in the AMA Manual of Style (Iverson et al., 2007, p. 519). Iverson et al. (2007,

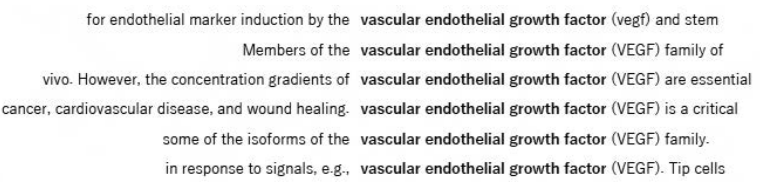


Figure 1: Example of concordance lines for *vascular endothelial growth factor*

p. 501) suggest that “Use common sense in deciding whether to abbreviate the terms.” In the DDL setting, instructors may guide the learners to become aware of the usage and form of such terms and their abbreviations.

The first letter of the phrase *In the study, we* was capitalized in all 145 sentences of the concordance (Figure 2). The phrase was followed by verbs describing actions such as *investigate* and *have developed* or reporting verbs such as *demonstrate*. In classrooms, learners can be guided to notice that the use of “this study” indicates the research being reported in the paper itself. They can also learn about the types and tense of verbs that follow the phrase *In this study, we*.

blood vessel growth, and cancer invasion.	**In this study, we** investigate the influence
of the proteins ERK1 and 2 (ERK1/2).	**In this study, we** have developed a
genetic determinants are largely unidentified.	**In this study, we** sought to determine
receptors and intracellular signaling pathways.	**In this study, we** generated an α5
(NFs) into CAFs is largely unknown.	**In this study, we** determined the contribution
from human monocytes. Methodology and Results:	**In this study, we** demonstrate the molecular

Figure 2: Example of concordance lines for *in this study, we*

Word profiles of the corpus suggested that the the CAPHYRA corpus might have lexical features indicative of medical RAs. The most frequent four-word expressions were classified into three types: highly specific technical expressions in the cardiovascular field, technical words of Latin origin, and hint expressions which were used for rhetorical purposes. The concordance lines with these expressions should be useful for activities in the classroom.

3. Examination of a larger corpus

3.1 Compilation of a medical research article corpus—The MEDRA corpus

The CAPHYRA corpus included only the RA abstracts of cardiovascular physiology papers. Building a larger corpus with AntCorGen was, therefore, studied from the viewpoint of pedagogical application.

Four corpora comprising 4,863 cardiology, 5,552 gastroenterology, 5,524 pulmonology, and 4,821 cancer RA texts were built successfully by the some procedure as that for the CAPHYRA corpus. The collection of cancer RA texts was limited to

"clinical oncology," "oncology agents," "cancer risk factors," and "cancer detection and diagnoses" to obtain approximately 5,000 RAs. The disciplines were selected based on consultation with an informant specializing in histopathology and gynecology.

From the four corpora, 100 RAs each were "randomly sampled" (L. Anthony, personal communication, June 13, 2017) to obtain a medical RA (MEDRA) corpus. The MEDRA corpus, consisting of 400 RA body texts from the four disciplines, was examined using AntConc (Anthony, 2014) and CasualConc (Imao, 2017).

3.2 Findings—Profiles of the MEDRA corpus

The corpus had 33,734 types and 1,778,417 tokens. Each portion of the four areas, including cardiology, gastroenterology, pulmonology, and cancer, had approximately 15,000 types and 450,000 tokens (Table 3). The 30 most frequent words in the MEDRA corpus and the normalized frequency (NF) are shown in Table 4. Ten texts each from the four portions were sampled and compared with 10 texts from all 400 texts. The 30 most significant words according to "log-likelihood ratio" (LLR; Dunning, 1993, p. 68) were identified using CasualConc (Imao, 2017) and shown in Table 5.[1] In the table, the words having "document frequency" (DF) of at least five (Tabata, 2012, p. 3) are in bold letters.[1]

In the MEDRA corpus as a whole (Table 4), the words *patients* and *cells* occurred 8,150 times (4.6 per 1000 words) and 7,578 times (4.3 per 1000 words), respectively, in 318 and 213 files. The letter *p* appeared times (5.0 per 1000 words) in 380 files and was mostly used to express *p* values.

(1) **Patients** with early-stage or minimal residual disease usually have lower levels of ctDNA, making it difficult to precisely detect specific alterations.
(Fan, Zhang, Yang, Ding, Wang, & Li, 2017, p. 2 [emphasis added])

(2) No single gene responsible for the commitment of mesenchymal **cells** to the angioblast cell fate has been identified as yet.
(Sumanas & Lin, 2006, p. 60 [emphasis added])

(3) In the combined group of stages II and III, serum mSST remained as a significant independent predictor of worse OS (HR = 2.797, 95% CI, 1.34–5.84; ***P*** = 0.006). (Fan et al., 2017, p. 12 [emphasis added])

Table 3: Word profiles of the MEDRA corpus as a whole and for each portion

	The MEDRA corpus as a whole	Cardiology portion	Pulmonology portion	Gastroenterology portion	Cancer portion
Type	49,542	15,656	16,588	17,298	14,666
Token	1,356,539	434,818	459,558	462,163	421,878

Table 4: The 30 most frequent words in the MEDRA corpus
(Normalized per 1000 words)

Rank	Word	NF	Rank	Word	NF	Rank	Word	NF	Rank	Word	NF	Rank	Word	NF
1	the	47.6	7	with	11.8	13	as	6.2	19	at	4.3	25	c	3.5
2	of	37.0	8	for	10.7	14	is	5.8	20	cells	4.3	26	study	3.3
3	and	30.9	9	were	10.4	15	or	5.3	21	on	4.3	27	not	3.3
4	in	26.4	10	was	9.6	16	p	5.0	22	are	3.6	28	s	3.2
5	to	17.1	11	that	6.6	17	from	4.8	23	we	3.5	29	be	3.1
6	a	15.8	12	by	6.5	18	patients	4.6	24	this	3.5	30	n	2.9

Table 5: The 30 most significant key words with ten texts from each portion
(Sorted according to LLR)

Rank	Cardiology			Pulmonology			Gastroenterology			Cancer		
	Word	DF	LLR	Word	DF	LLR	Word	DF	LLR	Word	DF	LLR
1	**artery**	**7**	**132.7**	infants	4	180.5	pain	3	124.0	**cancer**	**10**	**227.4**
2	pathways	1	105.5	neonatal	4	105.5	uc	2	117.1	oral	2	106.7
3	cc	2	93.9	deaths	4	103.0	asa	1	85.3	aging	2	81.4
4	vascular	3	62.4	**mortality**	**5**	**87.8**	liver	4	84.2	ga	1	73.7
5	air	2	59.0	**infection**	**7**	**83.1**	tnbs	1	70.0	**ci**	**7**	**71.3**
6	ci	4	54.0	mutation	2	76.4	**no**	**10**	**69.6**	occult	2	71.2
7	**cardiac**	**7**	**53.4**	cd117	2	74.1	mortality	3	57.8	**d**	**5**	**65.4**
8	**dysfunction**	**5**	**49.0**	resistance	4	58.6	coverage	2	57.7	relapse	2	62.2
9	**coronary**	**6**	**47.6**	**children**	**6**	**52.1**	duration	4	54.8	**vs**	**5**	**61.9**
10	**rate**	**8**	**47.2**	**respiratory**	**7**	**52.1**	us	3	47.3	screening	4	61.4
11	**m**	**6**	**42.8**	x	4	51.9	children	3	47.0	**growth**	**5**	**60.0**
12	**vs**	**8**	**41.3**	**subject**	**6**	**51.0**	colonic	3	46.2	**diagnosis**	**6**	**56.4**
13	access	3	39.5	**death**	**6**	**50.1**	lamina	2	40.7	prognostic	3	48.3
14	origin	1	39.3	**causes**	**5**	**40.8**	propria	2	40.7	kg	2	47.6
15	wild	2	36.7	mca	1	39.9	improved	3	39.6	**response**	**7**	**44.0**
16	**post**	**6**	**34.8**	**rates**	**6**	**39.9**	predictor	2	39.6	survival	3	43.2
17	**baseline**	**5**	**34.7**	**n**	**9**	**37.5**	epithelium	1	36.4	liver	3	41.7
18	yes	1	33.8	month	3	36.8	**severity**	**5**	**33.4**	**status**	**7**	**40.5**
19	**reported**	**10**	**33.4**	**kit**	**5**	**35.4**	costs	1	33.1	symptoms	3	39.3
20	**tests**	**5**	**32.4**	born	3	35.1	apoptosis	3	32.5	**et**	**8**	**37.9**
21	monocytes	1	31.4	individuals	3	30.3	genotype	2	31.4	apoptosis	4	36.9
22	san	3	31.4	**culture**	**6**	**30.0**	**model**	**7**	**30.2**	**al**	**8**	**36.9**
23	**classification**	**5**	**30.4**	**rate**	**6**	**27.9**	health	4	29.8	**reference**	**5**	**36.0**
24	associations	2	30.1	**incidence**	**5**	**26.1**	drug	1	28.5	score	4	34.0
25	**studies**	**9**	**29.6**	**due**	**9**	**23.3**	**effects**	**8**	**28.2**	**diagnosed**	**7**	**33.2**
26	**during**	**10**	**29.2**	delivery	2	23.2	**genes**	**5**	**26.8**	**patients**	**10**	**32.9**
27	**no**	**10**	**29.1**	**risk**	**6**	**23.0**	**tests**	**5**	**26.6**	months	4	32.9
28	constant	2	27.8	responses	3	20.9	**al**	**9**	**26.1**	**table**	**9**	**32.5**
29	course	4	27.8	hospitalization	3	20.4	penetration	1	24.8	**median**	**6**	**32.3**
30	**type**	**9**	**27.3**	cd44	1	20.0	processes	2	24.8	**baseline**	**5**	**32.0**

The DF values in Table 5 indicate that many of the most significant words occurred in less than half of the texts, suggesting that each portion of the MEDRA corpus might be diverse in their word profiles. Some of the words, including *ci*, *rate*, *vs*, *baseline*, *tests*, and *no*, occurred at least five times in the texts sampled from two different portions.

3.3 Classification of the MEDRA corpus by text-type

The examination suggested that each portion of the MEDRA corpus may have various types of texts. Swales (1990, p. 19) supports the idea of Geertz (1983), who indicated that "Grand rubrics like 'Natural Science' ... and 'The Humanities' ... merely block from view what is really going on out there." Salager-Meyer (2014, p. 50) regards genres as "text-types" and raises the examples of medical genres such as "research articles," "case reports," and "review articles." Williams (1996, p. 175) defines "two types of research articles" as "clinical" and "experimental" in his contextual study of verbs in medical RAs.

According to the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH), which is an international non-profit association for ensuring "development and registration of safe, effective, and high quality medicine" (ICH, *n.d.*), the regulatory format for application of new drugs has separate sections for clinical and non-clinical texts. The informant who participated in this study commented that "[o]ne of the major factors that influence RA writing may be the degree of homogeneity in the methods used in studies. Studies with human subjects may not be as homogeneously controlled as those using animals or cells. For example, a clinical study may enroll participants with different dietary habits. ... Therefore, the rationale of study design such as experimental systems and inclusion criteria should be discussed in RAs."

Based on these considerations, the RAs in the MEDRA corpus were classified into five text-types (National Center for Global Health and Medicine, 2009): in vitro studies, in vivo animal studies, reviews, case and cohort studies, and meta-analysis based on discussion with the informant.[2]

The corpus texts were classified by visiting the original RA websites, observing the hint expressions in the title, abstract, and the methods section, if necessary. The following text was classified as an in vivo animal study:

(4) **Weekly** Doxorubicin **Increases** Coronary Arteriolar Wall and Adventitial Thickness

... Doxorubicin (DOX) **is associated with** premature cardiovascular events **including** myocardial infarction. **This study was performed to determine if** the **weekly administration of** DOX **influenced** coronary arteriolar medial and/or adventitial wall thickening. ... Thirty-two **male** Sprague-Dawley **rats aged** 25.1 ± 2.4 **weeks were randomly divided into** three **groups and received weekly** intraperitoneal **injections of...high** (2.5 mg/kg, n=11) **doses of** DOX. (Eckman et al., 2013, p. 1 [emphasis added])

The hint expressions in the example text included *weekly administration of* and *male...rats,* and the text was classified as an in vivo animal study using rats. Of the 400 texts, 14 were classified as in vitro studies; 36 as in vivo animal studies, 204 as case and cohort studies, 23 as meta-analysis, and 10 as reviews.[3] Ten texts each from the five text-types were sampled for comparison with 10 samples of the 400 texts which were different from those used for the data in Table 5. Table 6 shows the 30 most significant words according to LLR obtained by CasualConc (Imao, 2017).[1] The words in bold letters occurred in at least five (50%) of the sampled texts.

3.4 Pedagogical implication

Concordance lines revealed that the word *cell* in the in vitro study portion was frequently used to modify other nouns in phrases such as ***cell*** *culture* and ***cell*** *line.* The word *cells* observed in the animal study portion was, however, used as a noun in phrases such as *cancer* ***cells*** and *endothelial* ***cells***. The word *ci* in the case cohort study portion appeared solely in the phrase *95%* ***CI*** and occurred ten times more than its expanded form *confidence intervals*. These features are noteworthy for pedagogy.

In classrooms, the MEDRA corpus could be used for learners from various disciplines in medicine. The concordance lines with a node word of learner's interest, such as "cells" could be presented, and the usage of the word in paragraphs or sections could be presented to identify the text-type or genre. Scott and Tribble (2006, p. 109) indicate that combinations of "KW [keyword] analysis and discourse analysis" are useful for understanding "how language is used." Further study is needed to explore more approaches for pedagogy.

Table 6: The 30 most significant key words with ten texts from five text-type portions (Sorted according to LLR; the MEDRA corpus)

Rank	In vitro			Animal			Case cohort			Meta analysis			Review		
	Word	DF	LLR	Word	DF	LLR	Word	DF	LLR	Word	DF	LLR	Word	DF	LL
1	**cells**	**9**	**390.4**	**mice**	**5**	**632.5**	**cases**	**5**	**145.1**	**survival**	**6**	**127.8**	**vs**	**8**	**291.8**
2	**expression**	**9**	**245.0**	**ca**	**7**	**196.5**	**cancer**	**5**	**110.5**	**pooled**	**7**	**92.7**	**months**	**6**	**169.4**
3	**mice**	**7**	**245.0**	**group**	**10**	**141.3**	crt	1	110.4	**meta**	**8**	**74.2**	**cohort**	**7**	**151.7**
4	**genes**	**6**	**206.8**	**rats**	**5**	**128.8**	**positive**	**6**	**104.2**	**analysis**	**9**	**69.2**	fc	1	129.4
5	monocytes	2	**171.8**	ko	1	114.3	df	1	101.2	**studies**	**10**	**65.5**	**studies**	**10**	**125.7**
6	radiation	3	**152.9**	**mouse**	**6**	**113.9**	hbv	1	99.4	surgery	3	58.3	water	3	104.1
7	**gene**	**6**	**125.6**	glucose	2	113.0	prevalence	4	96.0	**heterogeneity**	**8**	**57.2**	breast	4	94.8
8	**figure**	**8**	**118.8**	**b**	**10**	**110.2**	leakage	1	91.5	**methods**	**10**	**50.1**	activities	2	91.0
9	**cell**	**10**	**92.0**	**weight**	**9**	**109.8**	survival	4	77.5	**dose**	**7**	**48.6**	rt	1	84.0
10	**fold**	**6**	**90.2**	**induced**	**9**	**86.6**	activity	4	61.7	method	4	44.9	v	3	82.1
11	**protein**	**8**	**85.2**	**increased**	**10**	**84.0**	breast	1	57.0	hcc	1	42.6	**systematic**	**9**	**80.5**
12	**ca**	**8**	**78.8**	**mg**	**9**	**74.9**	metastasis	3	55.1	incidence	4	42.4	**cancer**	**7**	**78.4**
13	**liver**	**5**	**78.8**	**expression**	**8**	**74.5**	limitation	3	52.8	**subgroup**	**6**	**39.3**	**review**	**10**	**76.9**
14	**f**	**6**	**77.5**	**stimulation**	**5**	**73.8**	copd	2	51.3	af	1	37.7	**reported**	**10**	**69.5**
15	**human**	**8**	**68.2**	**changes**	**9**	**71.5**	lymph	3	48.9	**cancer**	**6**	**37.1**	**level**	**9**	**67.5**
16	**induced**	**8**	**62.3**	**normal**	**9**	**69.0**	nodes	2	48.3	supplementation	1	36.8	**study**	**10**	**66.8**
17	**controls**	**7**	**61.5**	**pressure**	**6**	**65.5**	**bmi**	**5**	**46.2**	**difference**	**8**	**36.0**	comorbidities	1	64.3
18	**c**	**10**	**60.6**	**body**	**8**	**63.8**	plasma	4	45.7	**effect**	**8**	**33.6**	**or**	**10**	**55.5**
19	nuclear	2	**60.3**	infected	2	63.3	**subjects**	**7**	**45.6**	estimated	4	31.4	publications	3	52.3
20	**using**	**10**	**60.2**	infection	3	63.3	infection	2	44.9	**disease**	**10**	**31.2**	**strength**	**5**	**50.6**
21	hcc	3	**52.9**	transmission	1	58.3	scan	3	44.6	nd	2	30.0	radiotherapy	3	49.0
22	infection	2	**51.8**	**significantly**	**9**	**56.6**	**variables**	**10**	**43.8**	yes	2	28.6	**regarding**	**8**	**47.3**
23	**significantly**	**10**	**48.8**	**figure**	**7**	**56.5**	**age**	**10**	**43.2**	**tumors**	**5**	**28.4**	**one**	**10**	**46.2**
24	**tumor**	**7**	**48.7**	**protein**	**7**	**55.8**	pressure	4	42.6	**plot**	**7**	**27.7**	alterations	2	44.1
25	μm	4	**48.5**	**c**	**10**	**53.9**	**index**	**5**	**40.4**	rt	2	27.7	**criteria**	**9**	**40.7**
26	**by**	**10**	**46.7**	**function**	**8**	**53.7**	wall	2	40.1	**forest**	**7**	**26.9**	**exclusion**	**7**	**39.0**
27	macrophages	1	46.2	transient	4	**52.9**	controls	3	39.1	restricted	2	26.6	indicator	3	38.9
28	**pathway**	**7**	**46.2**	**d**	**8**	**49.2**	females	4	37.3	**bias**	**8**	**25.6**	**described**	**10**	**37.7**
29	**change**	**8**	**45.7**	**weeks**	**8**	**48.0**	**cohort**	**7**	**37.1**	**included**	**10**	**25.4**	nr	2	36.9
30	acid	4	45.2	af	2	**47.6**	**factors**	**10**	**35.8**	oral	3	24.8	cc	2	34.2

4. Further research directions

The word profiles of the two corpora were identified in this study; the next step will be to use the corpora in writing classrooms. Handford (2013) argues that it is essential to understand "the context in which the text is produced, and the constraints under which the writer is working" (p. 260) and that "such understanding can then be transferred to academic writing requirements the students may have" (p. 261). Activities such as exploring lexico-grammatical patterns could be attempted. Concordance lines for *cancer* in MEDRA corpus, for example, show that it is frequently used to modify another noun, as in *a **cancer** patient*; it is often modified by other nouns, as in *breast **cancer*** and *colon **cancer***; and it is also modified using an adjective, as in *esophageal **cancer*** and *pancreatic **cancer***. This suggests patterned usage in the discourse community,

which is worth examining in class. Activities could also be attempted from the genre-based approach. The phrase *in this study, we* in the MEDRA corpus occurred in the discussion section of approximately 80 texts and was seen in the introduction section of about 40 texts. It should be suggested that "[s]tudents run a concordance" (Stevens, 1991, p. 45) on the phrase *in this study, we* in the corpus and learn the rhetorical patterns. The combination of the genre-based framework and corpus-based study may allow us to learn specific usages of words and phrases that help students develop their proficiency in discipline-specific writing.

5. Conclusion

The corpora were built successfully with AntCorGen. The CAPHYRA corpus was considered to have linguistic features of medical RAs of the field. The MEDRA corpus should be useful for pedagogical applications for students who are or will be conducting various types of studies in medicine.

Corpora built using AntCorGen, with file names enabling access to the original texts, can be used to identify high frequency phrases in the entire corpus or a portion of it to identify the communicative purposes (genre or text-types). Corpora can also be used to accumulate the knowledge that needs to be presented for the genre in classroom settings.

Acknowledgments

The author would like to express sincere gratitude to Tomoji Tabata, Hisashi Iwane, Yasuhiro Imao, Maki Miyake, Hodošček Bor, Nobuyuki Hino, Judy Noguchi, Tomoko Wakasa, and three anonymous reviewers for their invaluable comments on the manuscript. The author also wishes to thank the editor of this journal. The author is also grateful to audience at the 43rd Annual Conference of JAECS at Kwansei Gakuin University in 2017 for their helpful comments. Any remaining errors are the author's responsibility.

Notes

1. The data are available upon request.
2. Several types of studies (National Center for Global Health and Medicine, 2009) were

combined for categorizing purposes.

3. The remaining texts were classified as *Other* because they were unclassifiable, but they were included in the analysis.

References

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Anthony, L. (2017a). AntCorGen (Version 1.1.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Anthony, L. (2017b). Automating the construction of individualized discipline-specific corpora for data-driven learning. In I. Holliday, K. Hyland, & L. L. C. Wong (Eds.), *Conference Programme Book for the Center for Applied English Studies (CAES) International Conference: FACES of ENGLISH 2, Teaching and Researching Academic and Professional English* (p. 71). Hong Kong.

Caze, A. L. (2011) The role of basic science in evidence-based medicine. *Biology & Physiology, 26(1),* 81–98. DOI 10.1007/s10539–010–9231–5

Coates, R., Sturgeon, B., Bohannan, J., & Pasini, E. (2002). Language and publication in Cardio-vascular Research articles. *Cardiovascular Research, 53,* 279–285.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19(1),* 61–74.

Eckman, D. M., Stacey, R. B., Rowe, R., D'Agostino Jr., R, Kock, N. D., Sane, D. C., ... Hundley, W. G. (2013). Weekly doxorubicin increases coronary arteriolar wall and adventitial thickness. *PLOS ONE, 8(2)E57554,* 1–6.

Fan G, Zhang K, Yang X, Ding J, Wang Z, & Li, Z. (2017). Prognostic value of circulating tumor DNA in patients with colon cancer: Systematic review. *PLOS ONE, 12(2),* 1–17

Feak, C. B., & Swales, J M. (2012). *Academic Writing for Graduate Students, Essential Tasks and Skills* (3rd ed.). MI, USA: University of Michigan Press.

Flowerdew, J., & Li, Y. (2007). Language re-use among Chinese apprentice scientists writing. *Applied Linguistics, 28(3),* 440–465.

Flowerdew, J., & Wang, S. H. (2017). Teaching English for research publication purposes with a focus on genre, register, textual mentors and language re-use: a case study. In J. Flowerdew & T. Costley (Eds.): *Discipline-Specific Writing: Theory into Practice* (pp. 144–161). Oxon, UK: Routledge.

Geertz, C. (1983). *Local Knowledge: Further Essays in Interpretive Anthropology*. New York: Basic Books.

Handford, M. (2013). What can a corpus tell us about specialist genres? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 255–269). Oxford, UK: Routledge.

Hunston, S. (2013). How can a corpus be used to explore patterns? In A. O'Keeffe & M.

McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 152–166). Oxford, UK: Routledge.

ICH. (*n.d.*). M4: The Common Technical Document. URL: http://www.ich.org/products/ctd.html

Imao, Y. (2017). CasualConc (Version 2.0.7) [Computer Software]. Osaka, Japan: Osaka University. Available from https://sites.google.com/site/casualconc/

Iverson, C., Christiansen, S., Flanagin, A., & Fontanaroas, P. B. (2007). *American Medical Association Manual of Style* (10th ed.). New York: Oxford University Press.

Johns, T. (1991). From print out to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Johns & P. King (Eds.), *English Language Research Journal, 4,* 27–45.

Jones, M., & Dimant, P. (2013). What can a corpus tell us about vocabulary teaching materials? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 387–400). Oxford, UK: Routledge.

Lee, D. (2008). Corpora and discourse analysis: New ways of doing old things. In V. Bhatia, J. Flowerdew, & H. Jones (Eds.), *Advances in Discourse Studies* (pp. 86–99). London: Routledge.

Lee, D. & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes, 25,* 56–75.

Maher, J. (1986). The development of English as an international language of medicine. *Applied Linguistics, 7(2),* 206–220.

Marco, M. J. L. (2000). Collocational frameworks in medical research papers: a genre-based study. *English for Specific Purposes, 19,* 63–86.

Maswana, S., Kanamaru, T., & Tajino, A. (2015). Move analysis of research articles across five engineering fields: What they share and what they do not. *Ampersand, 2,* 1–11.

Mauranen, A., (2017). Academically speaking: English as the lingua franca. In I. Holliday, K. Hyland, & L. L. C. Wong (Eds.), *Conference Programme Book for the Center for Applied English Studies (CAES) International Conference: FACES of ENGLISH 2, Teaching and Researching Academic and Professional English* (p. 22). Hong Kong.

Mizumoto, A., Hamatani, S., & Imao, Y. (2017). Applying the bundle-move connection approach to the development of an online writing support tool for research articles. *Language Learning, 67(4),* 885–921.

Nakatani, Y., & Bucsis, C. (Ed.). (2016). *Daigakusei no Tame no Academic Eibun Writing* [Academic Writing Strategies for University Students]. Tokyo: Taishukan Publishing Co., Ltd.

National Center for Global Health and Medicine. (2009). *Shoki Rinsho de Minitsuketai Rinsho Kenkyu no Essence* [Essence for Early Phase Clinical Research]. URL: http://www.imcj-gdt.jp/topics_04.pdf

Nesi, H. (2013). ESP and corpus studies. In B. Paltridge & S. Starfield (Eds.), *The Handbook of English for Specific Purposes* (pp. 407–426). West Sussex, UK: Wiley Blackwell.

Noguchi, J., (2004). A genre analysis and mini-corpora approach to support professional writing by nonnative English speakers. *English Corpus Studies, 11*, 101–110.

Noguchi, J., Matsuura, K., & Haruta, S. (2015). *Judy Sensei no Eigo Kagaku Ronbun no Kakikata Zoho KaiteiBan* [An Efficient Approach to Writing Up Your Research]. Tokyo, Japan: Kodansha.

Nwogu, K. N. (1997). The medical research paper: Structure and functions. *English for Specific Purposes, 16(2)*, 119–138.

Paltridge, B. (2013). ESP and pedagogy. In B. Paltridge & S. Starfield (Eds.), *The Handbook of English for Specific Purposes* (pp. 347–366). West Sussex, UK: Wiley Blackwell.

Pocock, G., Richards, C. D., & Richards, D. A. (2006). *Human Physiology*. UK: Oxford University Press.

Salager-Meyer, F. (1989). Principal component analysis and medical English discourse: An investigation into genre analysis. *System, 17(1)*, 21–34.

Salager-Meyer, F. (2014). Origin and development of English for medical purposes. Part I: Research on written medical discourse. *Medical Writing, 23(1)*, 49–51.

Scott, M., & Tribble, C. (2006). *Textual Patterns*. Amsterdam: John Benjamins.

Stevens, V. (1991). Classroom concordancing: Vocabulary materials derived from relevant authentic text. *English for Specific Purposes, 10,* 35–46.

Sumanas, S., & Lin, S. (2006). Ets1-related protein is a key regulator of vasculogenesis in zebrafish. *PLOS Biology, 4(1)*, 0060–0069.

Swales, J. M. (1990) *Genre Analysis: English in Academic and Research Setting*. UK: Cambridge University Press.

Tabata, T. (2012). Dickens to Collins no kyocho sakuhin heno buntai tokeigakuteki approach. *IPSJ SIG Technical Report, 2012-CH-93(3),* Information Processing Society of Japan.

Tojo, K., Hayashi, H., & Noguchi J. (2014). Linguistic dimensions of hint expressions in science and engineering research presentations. *JACET International Convention Selected Papers. 1*, 131–163.

Williams, I. A. (1996). A contextual study of lexical verbs in two types of medical research report: Clinical and experimental. *English for Specific Purposes, (15)3*, 175–197.

（浅野　元子　大阪大学大学院　Email: u029294j@ecs.osaka-u.ac.jp）

「特別講演」

The ICNALE Edited Essays: A Dataset for Analysis of L2 English Learner Essays Based on a New Integrative Viewpoint

Shin'ichiro ISHIKAWA

1. An Outline of the ICNALE Project

The International Corpus Network of Asian Learners of English (ICNALE) is a large-scale collection of English speeches and essays produced by college students (including some graduate students) from ten countries and regions in Asia and native English speakers. The ICNALE currently consists of four data modules: Spoken Monologue (Ishikawa, 2014), Spoken Dialogue (Ishikawa, 2018), Written Essays (Ishikawa, 2013), and Edited Essays.

Table 1: Structure of the ICNALE

Module	Released	Samples	Tokens	Contents
Spoken Monologue	2016	4,400	c500,000	one-minute monologues recorded on the answering phone
Spoken Dialogue	2020	---	---	approximately 30-to-40-minute utterances in OPI-like interviews
Written Essays	2013	5,600	c1,300,000	200-to-300-word essays
Edited Essays	2018	640	c150,000	learners' original essays and their edited versions

The ICNALE has seven key principles: (1) a focus on Asia, (2) consideration of linguistic modes, (3) condition control, (4) proficiency control, (5) learner background survey, (6) native-speakers' reference data collection, and (7) open distribution.

First, the ICNALE focuses exclusively on Asian learners. Paying attention to the diversity of English learners/users in the region, it collects data in both EFL areas (China, Indonesia, Japan, Korea, Taiwan, and Thailand) and ESL areas (Hong Kong, Pakistan, the Philippines, and Singapore).

Second, the ICNALE collects varied modes of learner English: spoken and written, and also monologue and dialogue.

Third, the ICNALE controls the conditions for speaking and writing as rigidly as possible. The number of topics (prompts) has been restricted to two ("It is important for college students to have a part-time job" and "Smoking should be completely banned at all the restaurants in the country"). Participants are required to speak or write about what they think of the topics and why they think so. An essay is required to be between 200 words and 300 words in length, and the time allotted for a monologue speech is 60 seconds. As a rule, samples that have not met these criteria are excluded.

Fourth, the ICNALE collects L2 proficiency data from all the participants; in fact, they are required to report their scores in English proficiency tests such as TOEFL, TOEIC, and IELTS and also take a receptive vocabulary size test (Nation & Beglar, 2007). Thus, based on their scores in English proficiency tests or vocabulary size tests, all participants are classified into four proficiency bands linked to the Common European Framework of Reference for Languages (CEFR) scale: A2, B1−1 (B1 lower), B1−2 (B1 upper), and B2+.

Fifth, the ICNALE collects a wide range of background information from the participants, including sex, age, a period of time spent learning English, any experiences of staying in English speaking countries, motivations to learn English, language skills they like to focus on, experiences of using L2 at schools, and so on.

Sixth, the ICNALE also includes native-speakers' L1 English production data. They are given the same prompts and required to speak or write in the same situations. Considering possible diversity within the category of "native speakers" (Leech, 1998), the ICNALE collects data from three groups of native speakers: college students, English teachers, and other adults.

Finally, the ICNALE data is made available to researchers around the world. Users can download the whole ICNALE dataset and conduct their own research. In addition, they can access ICNALE resources through the online query system, which is called "The ICNALE *Online*."

The ICNALE project was launched in 2007. After ten years of continuing efforts by a group of international researchers, it has become one of the largest learner corpora ever built; it is now utilized by researchers, teachers, and students around the world.

2. An Outline of the ICNALE Edited Essays

2.1 Aim

The ICNALE Written Essays, which includes 5,600 essay samples, has been widely used since its release. However, the included essays are neither error-tagged nor rated, and a corpus user cannot discuss what type of errors tend to occur in learner essays, how the errors should be corrected, and what degree of quality the essays possess. In order to make this kind of deeper analysis of learner essays possible, we have released a new ICNALE module, the ICNALE Edited Essays. This module aims to become a reliable dataset that will enable the analysis of L2 English learner essays based on a new integrative viewpoint.

2.2 Contents

The ICNALE Edited Essays includes learners' original essays, their fully edited versions, and rubric-based evaluation scores; these features enable users to analyze the quality of the learner essays using an integrative viewpoint.

The original essays were chosen at random from the ICNALE Written Essays. Excluding several cases where the number of original essays was not sufficient, we took 20 essays written by learners at each of the four different proficiency levels (A2, B1−1, B1−2, and B2+).

Table 2: Number of samples in the ICNALE Edited Essays

		A2	B1−1	B1−2	B2+	Total
EFL	China	20	20	20	20	80
	Indonesia	20	20	20	NA	60
	Japan	20	20	20	20	80
	Korea	20	20	20	20	80
	Taiwan	20	20	20	20	80
	Thailand	20	20	20	NA	60
ESL	Hong Kong	NA	20	20	20	60
	Pakistan	NA	20	20	NA	40
	Philippines	NA	20	20	20	60
	Singapore	NA	NA	20	20	40
Total		120	180	200	140	640

2.3 Collection of essay evaluation data

Although there are varied approaches to essay evaluation, many of them involve using some kind of rating rubric, which helps raters to rate learner essays in a consistent and reliable manner. One of the most widely used rubrics in the field of TESOL is the ESL Composition Profile (Jacobs *et al.*, 1981), which uses five rating criteria: Content (CON), Organization (ORG), Vocabulary (VOC), Language use (LNU) , and Mechanics (MEC).

Therefore, we recruited five professional editors, all of whom are native English speakers with strong academic backgrounds and ample experience in editing academic papers for publication in major journals, and asked them to rate learner essays with reference to the ESL Composition Profile.

Table 3: Profiles of editors who participated in the ICNALE Edited Essays Project

Editors	Age	Sex	Degree	Years	L1 English
Editor A	28	Female	BA	3	Canadian
Editor B	32	Female	MS	5	Australian
Editor C	27	Female	BS	3	American
Editor D	38	Female	BS	10	British
Editor E	31	Female	PhD	2	Australian

Table 4: Scoring guide for the category of content

Score	Descriptors
10～12	EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic
7～9	GOOD TO AVERAGE: some knowledge of the subject • adequate range • limited development of thesis • mostly relevant to topic but lacks detail
4～6	FAIR TO POOR: limited knowledge of the subject • little substance • inadequate development of topic
1～3	VERY POOR: does not show knowledge of the subject • non-substantive • not pertinent • OR not enough to evaluate

In the original rubric, different scores were assigned to the five categories. However, we asked editors to score all the categories using 1–12 points so that they could rate more easily and consistently. Next, we calculated two kinds of total scores: a simple sum and a sum reflecting the weights suggested in the original rubric.

2.4 Collection of editing data

After rating the essays, the five editors were asked to edit learner essays on MS Word in track change mode so that clarity of the essays is improved and they become fully intelligible. They were required to retain the original texts as much as possible. Making any additions, deletions, and changes in the original contents was prohibited.

The figures below show how a learner's original essay was edited by an editor.

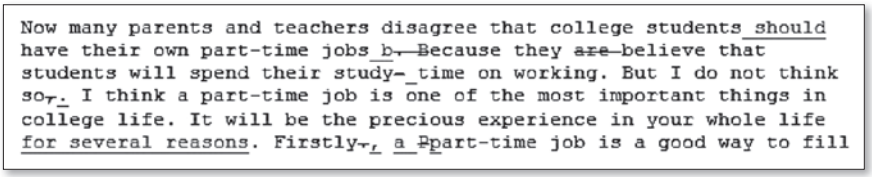

Now many parents and teachers disagree that college students should have their own part-time jobs b~~.~~ ~~Because they are believe~~ that students will spend their study~~-~~ time on working. But I do not think so~~.~~. I think a part-time job is one of the most important things in college life. It will be the precious experience in your whole life for several reasons. Firstly~~,~~, a ~~P~~part-time job is a good way to fill

Fig. 1: Sample of the edited essays

Fig. 2: Summary of changes added by an editor

As all the changes have been tracked in MS Word, corpus users can easily count how many words are added and deleted by an editor. In the sample above, 34 words were added, and 29 words were deleted. The number of edits, which is often called "an edit distance," is 63. Typically, the number of edits tends to decrease for good essays and increase for problematic essays.

2.5 Inter-editor variances

When a group of editors is in charge of scoring and editing, we need to be careful about possible inter-editor variances. Therefore, prior to the collection of evaluation and editing data, we conducted a small-scale calibration study in which we asked all five editors to rate and edit the same set of eight essays. These essays were written by learners who were at different proficiency levels and had different L1 backgrounds. Then, we confirmed the extent to which the average rating scores and the average numbers of edits could be maintained at stable levels across five editors.

Table 5: Average scores for eight samples

	Rating Scores (/12)						Number of Edits
	CON	ORG	VOC	LNU	MEC	Average	
Editor A	7.88	7.88	7.25	7.25	8.38	7.73	59.63
Editor B	9.88	10.38	8.38	7.63	8.25	8.90	49.50
Editor C	9.13	8.88	8.25	8.00	8.50	8.55	41.88
Editor D	8.38	8.13	7.88	7.25	8.00	7.93	48.50
Editor E	7.50	8.00	8.25	7.75	8.38	7.98	40.00
Average	8.55	8.65	8.00	7.58	8.30	8.22	47.90
Dif	2.38	2.50	1.13	0.75	0.50	1.18	19.63
Dif / Average (%)	27.78	28.90	14.06	9.90	6.02	14.30	40.97

The averages of the five category scores ranged between 7.73 (Editor A) and 8.90 (Editor B), and the gap was 1.18, which is equivalent to 14.30% of their average (8.22) and 9.79% of the full score (12). We could say that the variance is generally small in terms of essay evaluation. Meanwhile, the averages of the numbers of edits (*i.e.*, additions + deletions) ranged between 40.00 (Editor E) and 59.63 (Editor A), and the gap was 19.63, which amounts to 40.97% of their average (47.90). The variance was larger in terms of essay editing, though excluding Editor A, who tended to underrate learner essays and make more edits to them, the gap shrank to 21.13%.

This suggests that the rubric-based rating was more stable than editing, which required editors to define "intelligibility" by their own standards. The ICNALE Edited Essays includes detailed data of this calibration study and thus enables users to interpret the results of the data analysis in a more cautious and in-depth way.

2.6 Online query system

Users can download the entire dataset of the ICNALE Edited Essays and analyze it using any concordancer of their choice. In addition, they can access the data through the ICNALE *Online* interface, which currently offers two kinds of searches: KWIC Search and Keyword Search.

2.6.1 KWIC Search

KWIC Search enables users to retrieve the concordance lines including the target word(s). The figures below illustrate how users can examine the use of the term "go," which occurs in the original essays written by Chinese learners.

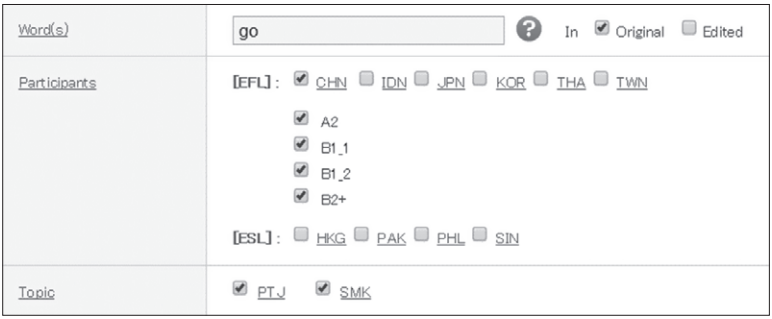


Fig. 3: Settings for the KWIC Search

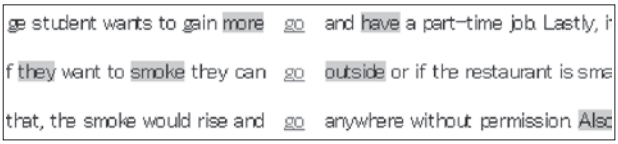


Fig. 4: Results of the KWIC Search

By entering a word or an expression that they would like to analyze, and choosing the version of the essays (original or edited), learners' countries/areas, their L2 proficiency levels, and the topic of the essays ("a part-time job" or "non-smoking"), users can obtain a list of concordance lines including the target word(s).

In the KWIC concordance, some words appear shaded; this shows that the words or phrases including them are changed in the edited essays. Fig. 4 shows that words such as "more," "have," "they," "smoke," "outside," and "Also" or expressions including them are to be altered in the edited versions.

When any word in the KWIC concordance is clicked, a new window pops up, and users are able to examine an original essay and its edited version at the same time.

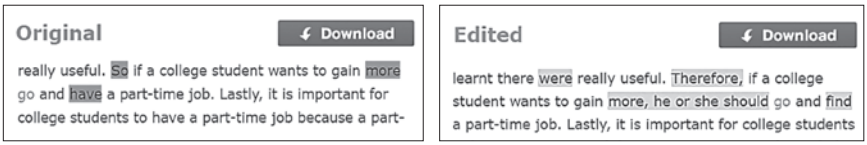


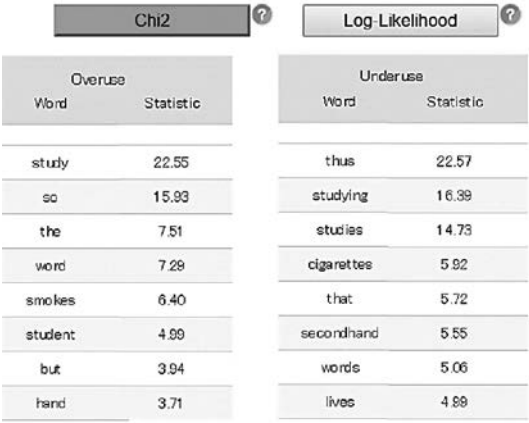
Fig: 5 Side-by-side comparison of the original/edited essays

Thus, users can easily analyze the kinds of problems that exist in learners' original essays and the correction of these problems by professional editors.

2.6.2 Keyword Search

Keyword Search enables users to identify words that occur at a statistically higher rate in one of the two texts to be compared.

Fig. 6: Settings for the Keyword Search



Overuse	
Word	Statistic
study	22.55
so	15.93
the	7.51
word	7.29
smokes	6.40
student	4.99
but	3.94
hand	3.71

Underuse	
Word	Statistic
thus	22.57
studying	16.39
studies	14.73
cigarettes	5.92
that	5.72
secondhand	5.55
words	5.06
lives	4.89

Fig. 7: Results of the Keyword Search

By choosing the target (original essays) and reference (edited essays) datasets, which need to be for the same learner group, users can obtain a list of words that were overused in the original essays (namely, words to be deleted in the edited essays) and words that were underused in the original essays (namely, words added in the edited essays). Users can choose Chi-square value or Log-likelihood value as a statistical method for the keyness calculation.

Fig. 7 suggests that learners often use “so” in an erroneous way and that it should be replaced by “thus,” for example. Thus, users can easily see the kinds of errors and inappropriate vocabulary that existed in learners’ original essays and how they were corrected.

3. A Case Study: Multivariate Analysis of Chinese Learners’ Essays

3.1 Aim and RQ

The ICNALE Edited Essays makes it possible to discuss the quality of learner essays based on a new integrative viewpoint. Using a part of its data, Ishikawa (In press) discussed the quality of Japanese learners’ essays, paying attention to the rubric-based rating scores, the number of edits made by editors, and the average keyness values showing the degree of inappropriate vocabulary use by learners in comparison to native English speakers.

In the current study, then, we will discuss the features of Chinese learners’ essays by considering essay-related parameters and writer-related parameters in an integrative way. Our research questions are as follows:

RQ1 To what extent are essay evaluation scores and the number of edits correlated?

RQ2 How are essay-related and writer-related parameters clustered?

3.2 Data and method

We analyzed 80 essays written by Chinese learners, which had been included in the ICNALE Edited Essays. Considering RQ1, we paid attention to the strength of the correlation between the sum of the five category-based evaluation scores and the number of edits made by editors.

Next, considering RQ2, we paid attention to the wide variety of essay-related and writer-related parameters shown below:

Table 6: Parameters used for the analysis

Essay-related parameters
Essay topics (part-time job [PTJ] or non-smoking [SMK]), Essay length (number of words per essay [LEN] ranging between 200–300 words), Number of edits (inverse of the number of additions and deletions [EDT]), Essay evaluation scores (five category-based scores ranging between 1–12: Content [CON], Organization [ORG], Vocabulary use [VOC], Language Use [LNU], and Mechanics [MEC], as well as their simple sum [TTL])
Writer-related parameters
Sex (female [FEM] or male [MAL]), Age (ranging between 18–21 years old [AGE]), Major (humanities [HUM], social sciences [SCS], or science and technology [SCT]), vocabulary size (scores in the receptive vocabulary size test (Nation & Beglar, 2007), ranging between 0–50 [VST]), Strength of motivation (integrative motivation [INT] and instrumental motivation [INS] ranging between 0–6 and both types [MOT] ranging between 0–12), Amount of L2 use (at primary schools [PRM], secondary schools [SEC], and colleges [COL], as well as in classes [INC] and out of classes [OTC] ranging between 1–6), Former L2 teaching (frequency of instruction by native English speaking teachers [NST] and specific instructions focusing on pronunciations [PRN], presentations [PRS], and essay writing [ESW], the type of L2 skills that learners like to focus on: (listening [LNS], reading [RDS], speaking [SPS], and writing [WRS], ranging between 1–6).

Discussing the number of edits, we used the inverse number, with the assumption that good essays would require fewer edits. Writer-related data, except for the vocabulary size, was obtained from the questionnaire survey.

We conducted a hierarchical cluster analysis in order to observe the interrelation between the variables. The distance was defined as the square root of (2−2*r*), and the Ward method was adopted for calculation of distance.

3.3 Results

3.3.1 RQ1 Strength of correlation

The table below presents Pearson's *r* values between the essay evaluation scores and the number of edits.

It was found that the number of edits showed a middle-level correlation (0.391–0.492) with any of the five category-based evaluation scores as well as their total score.

Table 7: Correlations between the essay evaluation scores and the number of edits

	EDT	CON	ORG	VOC	LNU	MEC	TTL
EDT	1.000						
CON	0.391	1.000					
ORG	0.436	0.786	1.000				
VOC	0.470	0.655	0.684	1.000			
LNU	0.443	0.644	0.684	0.742	1.000		
MEC	0.448	0.602	0.615	0.550	0.635	1.000	
TTL	0.492	0.904	0.896	0.841	0.865	0.716	1.000

This result seems to corroborate our common belief that good essays receive fewer edits and that bad essays receive more edits. However, it is important to note that the correlation *r* was not as high as was generally expected, which suggests that some of the Chinese learners wrote good but problematic essays, and others wrote poor but not so problematic essays. These findings show that grammatical and lexical correctness may not always guarantee the quality of learner essays.

3.3.2 RQ2 Clustering of essay-related and writer-related variables

By conducting a cluster analysis, we obtained a tree diagram, which is provided below. By setting a cutting point of 2, 33 variables were classified into five clusters (A, B, C, D, and E).

First, we would like to pay attention to Cluster E, which demonstrates that the essay evaluation scores and the number of edits are closely related. It also suggests that the relationship between evaluation and editing becomes clearer in essays about a social topic (non-smoking) rather than in essays about a familiar personal topic (a part-time job). Thus, we may infer that the essay topic may influence the essay's evaluation and its editing.

Next, Cluster B shows that the male students, many of whom have majored in science and technology, tend to have more experiences of being taught pronunciation, presentation, and essay writing, and they usually know more words in the target language. The experience of learning how to write good essays and the wider vocabulary size are expected to contribute directly to the quality of essays. However, as suggested by the fact that Clusters B and E do not merge until the distance reaches 2.5, such a connection is not necessarily confirmed by the current data.

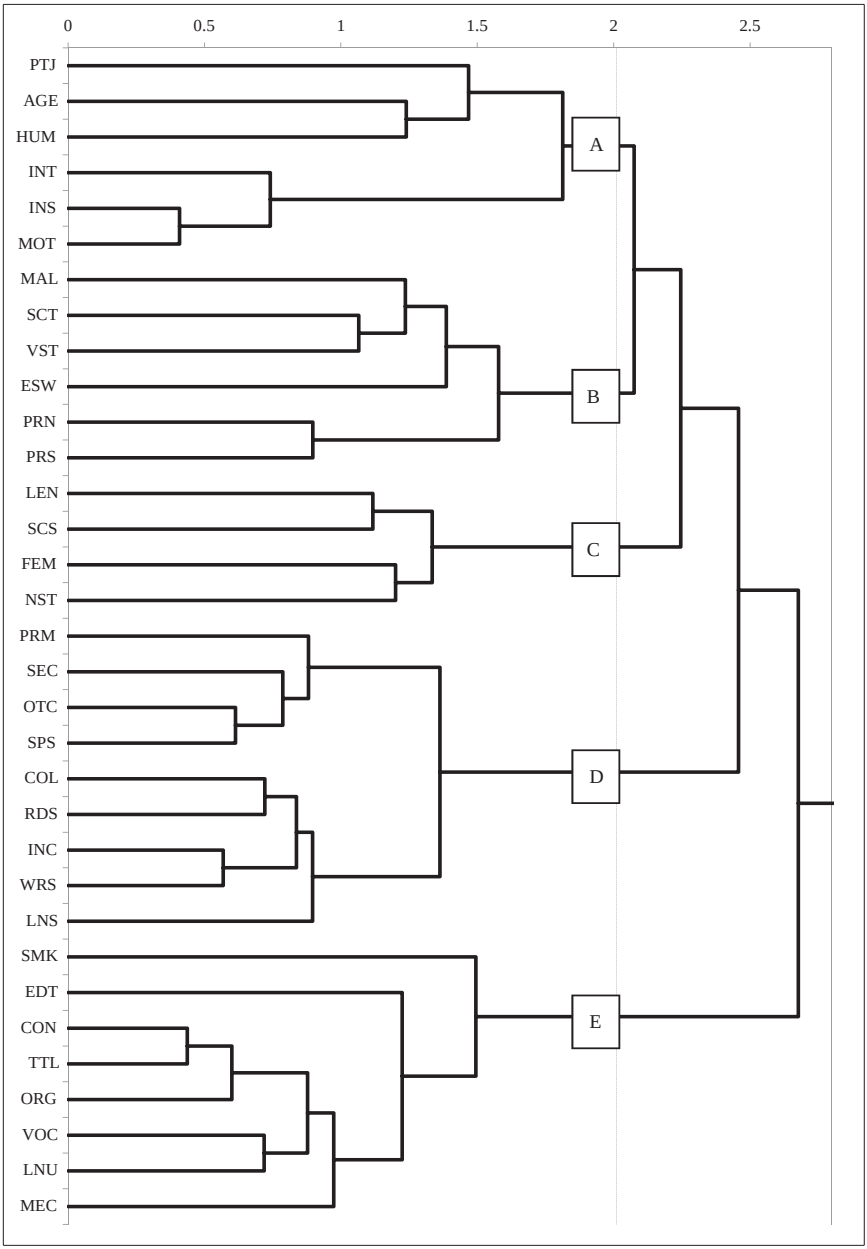


Fig. 8: A tree diagram based on the cluster analysis of the 33 variables

Table 8: The variables in each cluster

Cluster A	Cluster B	Cluster C	Cluster D	Cluster E
PTJ	MAL	LEN	PRM	SMK
AGE	SCT	SCS	SEC	EDT
HUM	VST	FEM	OTC	CON
INT	ESW	NST	SPS	TTL
INS	PRN		COL	ORG
MOT	PRS		RDS	VOC
			INC	LNU
			WRS	MEC
			LNS	

Other clusters also suggest interesting facts about the Chinese writers and their L2 essays. Cluster A shows that humanities students, including some English-major students, often have a higher L2 learning motivation and that they perform better when writing about a personal topic (a part-time job). Cluster C reveals that female students, many of whom majored in social sciences, tend to have more experiences of being taught by native English-speaking teachers, and they tend to write somewhat longer essays. Finally, Cluster D shows that learners' former experience of L2 use and the type of skills that they like to focus on are related to some extent. It is noteworthy that which skill learners have focused on in L2 learning mighr influence how much they have actually used L2 in various situations.

4. Summary

This paper introduced the outline of the ICNALE project and explained the aim, design, and contents of the ICNALE Edited Essays as its newest addition.

It then illustrated how the ICNALE Edited Essays could be used for the analysis of learner essays based on a new integrative viewpoint. Our case study, though quite preliminary, has proven that the essay evaluation scores and the number of edits show a middle-level correlation and that they can be influenced by the essay topic and also by a variety of essay-related and writer-related parameters. The author hopes that the ICNALE Edited Essays will contribute to spreading a new data-based analysis of the quality of learner essays, which is based on an integrative observation of texts and text producers.

Bibliography

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, *29* (3), 371–383.

Brooks, G. (2012). Assessment and academic writing: A look at the use of rubrics in the second language writing classroom. *Humanities Review* (Kwansei Gakuin University), *17*, 227–240.

Ishikawa, S. (2011). A new horizon in learner corpus studies: The aim of the ICNALE project. In G. Weir, S. Ishikawa, & K. Poonpon (Eds.), *Corpora and language technologies in teaching, learning and research* (pp.3–11). Glasgow, UK: University of Strathclyde Publishing.

Ishikawa, S. (2012). *Beshikku kopasu gengogaku*. Tokyo: Hitsuji Shobo. [A Basic Guide to Corpus Linguistics].

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

Ishikawa, S. (2014). Design of the ICNALE-Spoken: A new database for multi-modal contrastive interlanguage analysis. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 2*（pp. 63–76). Kobe, Japan: Kobe University.

Ishikawa, S. (2018). The ICNALE Spoken Dialogue no sekkei: Taiwa ni okeru L2 koto sanshutsu kenkyu no tame ni. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 3*（pp. 3–21). Kobe, Japan: Kobe University. [Design of the ICNALE Spoken Dialogue: For studies of L2 oral production in dialogues]

Ishikawa, S. (In press). Comparison of three kinds of alternative essay-rating methods to the ESL Composition Profile: An approach to lessen teachers' workloads in evaluating learners' L2 English essays.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

Leech, G. (1998). Preface. In S. Granger (Ed.), *Learner English on computer* (pp. xiv-xx). Harlow, England: Addison Wesley Longman.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.

（石川慎一郎　神戸大学　Email: iskwshin@gmail.com）

「シンポジウム」話し言葉コーパスの構築と利用

迫田久美子・野口ジュディー・長谷部陽一郎

まえがき

Abstract

This paper is a summary of a symposium on the Construction and Applications of Spoken Corpora, which was presented at the 43rd Annual Conference of the Japan Association of English Corpus Studies on October 1, 2017 at Kwansei Gakuin University, Nishinomiya, Hyogo, Japan. Four corpora and their interfaces were introduced: The ICNALE (The International Corpus Network of Asian Learners of English) by Shin'ichiro Ishikawa, The I-JAS (International Corpus of Japanese as a Second Language) by Kumiko Sakoda, JECPRESE (Japanese-English Corpus of Presentations in Science and Engineering) by Judy Noguchi and TED Corpus Search Engine by Yoichiro Hasebe. This paper covers the latter three as The ICNALE is described in a lecture paper in this issue. The background and rationale for the construction of the corpus or the software interface is introduced together with how they can be used for the benefit of language learning and teaching.

要旨

英語コーパス学会第 43 回大会（2017 年 10 月 1 日，関西学院大学）で開催されたシンポジウム「話し言葉コーパスの構築と利用」の内容を紹介する。シンポジウムでは，学習者の書き言葉・話し言葉（英語）ICNALE（石川慎一郎），学習者話し言葉日本語（テーマ別）I-JAS（迫田久美子），話し言葉日英（理系プレゼン）JECPRESE（野口ジュディー），TED コーパス検索システム（長谷部陽一郎） について取り上げられた。学習者の書き言葉・話し言葉（英語）ICNALE に関しては，当号の講演の論文に掲載している。本稿では，I-JAS と JECPRESS ならびに TED コーパス検索システムが構築された背景を説明し，言語教育への応用について紹介する。

『英語コーパス研究』第 25 号（2018），pp. 131–132

はじめに

野口ジュディー

言語学の多くの研究では書き言葉を対象とするが，人間の自然言語は，元来，話し言葉からスタートする。しかし，話し言葉は，研究の対象としては扱いにくいものである。コーパスを作成するためには，発話者から許可を得ないと録音することすら難しい。また，生きた言葉を捉える環境は，録音を行うのに適していないことが多い。このようなハードルを越えて録音ができたとしても，時間をかけて書き起こしを行う必要がある。最終的には，話し言葉コーパスを利用しやすくするために，インタフェースとなるシステムを用意しなければならない。このような作業の苦労話を研究者たちは少なくともいくつか抱えている。ここでは，研究者たちの努力によって構築された異なる４つのコーパスやコーパス検索システムを紹介する。なお，シンポジウムで紹介された ICNALE は，石川慎一郎氏による講演論文 “A Frontier in Learner Corpus Studies: For Better Understanding of L2 Learners” に詳しく紹介されている。以下には，学習者話し言葉日本語（テーマ別）I-JAS，話し言葉日英（理系プレゼン）JECPRESE，TED コーパス検索システムを，それぞれのパネリストが紹介する。

（神戸学院大学名誉教授　Email: jnoguchi@gc.kobegakuin.ac.jp）

「シンポジウム」
International Corpus of Japanese as a Second Language（I-JAS）：日本語学習者の言語研究と指導のために

迫田久美子・細井　陽子

1. はじめに

本稿は，現在構築中の日本語学習者コーパス（I-JAS）の概要を述べるとともに，学習者コーパスの重要性とそれを利用した研究を示すことを目的とする。「両親はふるさとで*（→に）住んでいます」「火の上に*（→で）魚を焼きます」など，日本語学習者には，格助詞の誤用がよく見られる（*は誤用を表す）。言語研究で対照分析研究が盛んだった 1950 ～ 60 年代，誤用は母語の干渉で起きると考えられ，習慣形成による文型練習が盛んにおこなわれた。しかし，異なった母語の学習者から同種の誤用が出現したり，母語転移の誤用の予測が外れたりしたことで研究の流れは対照分析から誤用分析に移っていった。

先述の「に」「で」の誤用は，母語に助詞のない中国語や英語話者にも，助詞を持つ韓国語話者の日本語学習者にも出現する。迫田（2001, 2002）は，先行研究や自身のコーパス（C-JAS）のデータから，学習者の「に」「で」の誤用に母語の影響よりは，学習者特有のストラテジーが影響していることを示した。初中級レベルの日本語学習者は，母語の違いにかかわらず「中・前・上」などの位置を表す語の後には「に」を，「会館・東京・大学」などの建物や地名を表す語の後には「で」を選択しやすく，「中に」や「会館で」のような固まりで覚えていると主張した。つまり，学習者は教師が教える「に」「で」の助詞の機能や分類で考えるよりも固まり（チャンク）を作って単語の一部として覚える可能性があることを指摘している。この考え方の出発点となったのは，学習者の多くの誤用例であった。

本稿は，学習者コーパスの重要性を述べ，現在構築中の日本語学習者のコーパス（I-JAS）の概要を示し，その I-JAS のデータから JFL（海外で日本語を学ぶ場合）と JSL（国内で日本語を学ぶ場合）の学習者における日本語習得の学習環境の影響について検討することを目的とする。本稿は，I-JAS 構築の統括責任者である迫田が 1. ～ 3. の執筆を担当し，JFL と JSL の学習者における日

『英語コーパス研究』第 25 号（2018），pp. 133–149

本語習得の学習環境の影響に関する内容は，迫田と細井の共同研究であるため，4. と 5. については共同で執筆を担当した。

2. 学習者コーパスの重要性

先述の「に」「で」の誤用は，いくつかの文献から収集されている。迫田（2001）は，文献の「寺」（寺村 1990），「福」（福間 1997），「市」（市川 1997），「迫」（迫田 1998）から，「に」「で」の誤用と正用例を抽出し，表 1 にまとめた。

表 1 「に」と「で」を含む誤用（●）と正用（○）（迫田 2001：19 を参考に）

句＼文献		寺	福	市	迫	誤用例（出典の文献）
「に」の誤用	後ろに		●		○	テレビの**うしろに＊**窓です。（福）
	中に	●	●	●	●○	その**中に＊**印象的だったのは〜。（市）
	前に	●		●		たばこ屋の**前に＊**会う〜。（市）
	隣に		●			〜**隣に＊**ピーターさんの部屋です。（福）
	〈地名〉に	●	●			**ネパールに＊**は今も 90％の人口は農業（福）
	田舎に					
	食堂に					
	大学に	●			●○	**大学に＊**試験を受けて落ちました。（迫）
「で」の誤用	後ろで					
	中で					
	前で					
	隣で		●			私は〜の**隣で＊**座っていました。（福）
	〈地名〉で	●	●	●	●	**東京で＊**住んでいます。（迫）
	田舎で			●	●	**タイの田舎で＊**は病院があるんで〜（市）
	食堂で			●	○	**食堂で＊**ごはんを食べに行きます。（市）
	大学で	●	●	●		ヘブライ**大学で＊**留学して，来年〜（寺）

迫田（2001）は，表 1 から「に」の誤用の前接名詞には位置を示す「中・前」が多く，「で」の誤用には建物や地名を示す単語が多い傾向があることに気づいた。そこで，日本語学習者は場所を表す「に」「で」を使用する場合に，次のような独自のルールを用いているのではないかと推測した。

（1）　位置を示す語（「中・前」ほか）＋「に」　　例　駅の前に
（2）　建物や地名を示す語（「会館」「東京」）＋「で」　例　国際交流会館で

迫田（2001）は，この仮説を中国語話者，韓国語話者，その他の言語を母語とする日本語学習者，および日本語母語話者の各 20 名ずつを対象として穴埋めテストを行った。その結果，図 1 が示すように初中級レベルの学習者は，母語の違いにかかわらず，「地名＋で」「位置＋に」の組み合わせでの正答率が高く，その反対は低いことがわかり，「に」「で」の使用は，前接する名詞と固まりで覚えている可能性が高いことを示した。

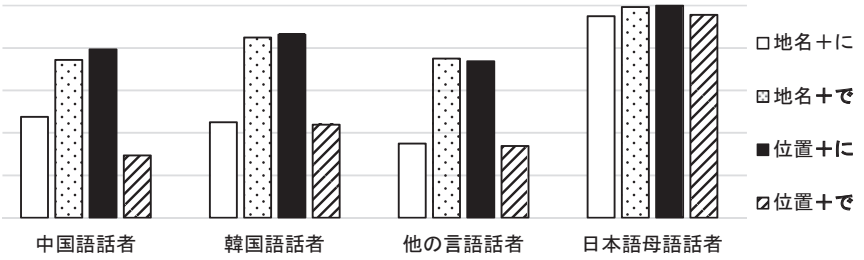


図 1　「に」「で」の調査における正答率（迫田2001：20）

この研究から，研究における重要な点を示すことができる。たとえば，外国語学習者は教師が教えたとおりの規則を使っているとは限らないこと，規範的な文法ではなく，固まりで覚えている場合も多いことなどである。しかし，最も重要な点は，日本語学習者の多くの誤用例に出会わなければ，上記の推測は生まれなかったことである。言い換えれば，多くの誤用例を見ることで学習者の独自のルールが見えてくるのである。

このように，外国語や第二言語の習得研究には，データが不可欠である。学習者のデータを見ずに，日本語学の観点から格助詞「に」「で」の用法の分類に沿って「に」「で」の使い分けを考え，以下のような多肢選択や穴埋めのテスト問題を作っても，学習者独自のルールは見えてこない。

（3）　もうすぐ行きますから，外（に・で）待っていてください。
（4）　傘は昨日使ったので，外（に・で）置いてあります。

Bley-Vroman（1983）は，学習者の誤用を論じる場合に母語話者の規範文法の基準や観点で分析したり，選択や穴埋め調査をしたりすることの危険性を「比較の誤謬（comparative fallacy）」と呼んでいる。また，野田他（2001）も，（5）のような「は」「が」の穴埋め調査だけでは，学習者の「は」「が」の習得を論じることは難しいと述べている。（* は，誤用を表す。）

（5）わたし（は*）遊んでいるとき，山田さん（が*）勉強しています。
　　→わたしが遊んでいるとき，山田さんは勉強しています。（野田他 2001：129）

学習者の習得プロセスや学習者独自の文法（中間言語）を探るためには，学習者自身の発話や作文を丹念に分析し，そこから学習者独自の見方，考え方を見つけ出すことが必要である。そのためにも，質的にも量的にも多種多様なコーパスが求められる。

3．学習者コーパス I-JAS

日本語学習者の習得研究や教育研究のニーズを背景に，1970 年代からさまざまな作文や発話コーパスが作られてきている。コーパスとは，作文や発話を大量に集めて電子化したデータベースの言語資料であり，母語話者汎用コーパスでは，1 億語から成るイギリス英語の BNC（British National Corpus）や日本語の BCCWJ（Balanced Corpus of Contemporary Written Japanese）が有名である。

英語学習者コーパスでも，今回のシンポジウムで取りあげられたコーパス（ICNALE, JECPRESE, TCSE）のみならず，さまざまなコーパスが作られ，データを基盤にした研究が進められている。

日本語学習者のコーパスでも 90 名の日本語学習者の発話データを所有する KY コーパスなどが作られ，日本語の習得研究に利用されている。しかしながら，多くの既存の日本語学習者コーパスは，英語学習者コーパスに比べると，学習者の母語も韓国語，英語，中国語が中心で母語の数が少なく，データ量においてもサイズが小さい。また，このことに加えて，既存のコーパスには，い

くつかの問題点がある。

1つには，学習者のレベルが明確でない場合も多い。滞日年数や学習期間を提示していても，必ずしもレベルを反映しているとは言えない。そのため，学習者間のレベルを比較することが難しい。

2つ目には，発話コーパスの場合，多くが対話のみであり，独話やさまざまなコミュニケーション場面のデータに乏しい。

3つ目には，学習者の情報が不足している。結果を考察する際に，多言語話者かどうか，家族内での言語，日本語によるアルバイト経験などの学習者の言語や学習環境の情報が重要となるが，学習者の背景情報を付与している既存のコーパスは極めて少ない。

このような問題点を考慮して構築している日本語学習者コーパスが「多言語母語の日本語学習者横断コーパス（International Corpus of Japanese as a Second Language: I-JAS）」である。本節では，I-JAS の特徴について述べる。

3.1　学習者の母語と学習環境

I-JAS の最も大きな特徴は，海外の日本語学習者の母語が 12 言語にわたっていることとコーパス完成時には，学習者のデータ 1,000 名，日本語母語話者のデータ 50 名の日本語学習者最大のコーパスとなることである。海外の学習者の 12 言語とは，言語類型論的な観点から選出した英語・中国語・韓国語・インドネシア語・スペイン語・フランス語・ドイツ語・タイ語・トルコ語・ハンガリー語・ベトナム語・ロシア語である。さらに，国内の学習者の場合は，海外のように言語の統制は取れていないが，日本語学校や大学等の教育機関で学んでいる教室環境学習者と就労者や外国人花嫁等の主に自然環境によって日本語を習得している自然環境学習者を対象とした。

海外での調査は 17 カ国 20 地域におよび，国内も東京，静岡，広島など複数の地域や教育機関で実施した。

このコーパスでは，母語の違いによる習得への影響を分析することができると同時に，環境の違いによる習得への影響も検討することができる。

3.2　タスクの多様性

次の特徴は，データのバリエーションが多いことである。このコーパスには，（1）日本語母語話者との対話（約 30 分）だけでなく，（2）独話で絵を見て物語を構築する「ストーリーテリング」2 種，（3）コミュニケーション場面を設

定した「依頼」と「断り」のロールプレイ 2 種,（4）1 枚の絵を描写するタスク,（5）スピーキングで用いたストーリーテリングと同じ題材のライティング,（6）テーマを与えられた作文データ（エッセイとメール）が揃っており，1 人の学習者の多様な言語バリエーションを検討することができる。タスクの詳細については，迫田他（2016）を参照してほしい。

これだけの多様なタスクを計画した背景には，できるだけ多くの研究テーマを網羅したデータを収集し，多くの研究で利用できるコーパスを作りたいという意図がある。対話は構造化インタビューで，Oral Proficiency Interview（OPI）の手法を取り入れ，一問一答の会話にならず，できるだけ自然な会話のやりとりを心掛けた。

また，同じストーリーテリングの題材で最初にスピーキングの調査をし，対面調査の最後に同じ題材でライティングを行った。そのため，同じ学習者のスピーキングとライティングの中間言語のバリエーションを分析することができる。

また，有志のみであるが，エッセイやメールの回答文の作文データが収集されているので，これも同様の研究のデータとして利用可能である。

3.3　統一されたレベル判定

I-JAS は，Japanese Computerized Adaptive Test（J-CAT）と Simple Performance Oriented Test（SPOT）という 2 つの客観テストを採用し，各学習者の点数をデータ化し，公開している。J-CAT は，聴解，文字・語彙，文法，読解の 4 つのテスト項目に分かれており，全体で 60〜90 分の時間を要する。日本語母語話者による対面の発話調査と PC による日本語能力テストは，互いに影響を受けることを避けるため，別々の場所で実施した。SPOT は，Tsukuba Test Battery of Japanese（TTBJ）の中の 1 つである。聞こえてくる文が画面に示され，その中の一字が抜けており，その一字を四択から選ぶ 90 問の問題である。1 問は 3 秒程度で処理されるため，約 5 〜 10 分で終了する。

3.4　学習者の背景情報

従来のコーパスには見られない特徴として，I-JAS には，学習環境や背景情報を詳細に公開している点がある。学習者の属性，現在の言語環境，学習歴，家族の母語，渡日経験，学習スタイル，日本語によるアルバイト経験，日本人の友人の有無など，研究倫理にも配慮しながら，対面調査に先立って 20 項目のアンケートをウェブ入力形式で行った。アンケートの言語は学習者の母語で

回答できるように 10 種類の言語で用意し，不明な点については，調査時に確認している。この情報により，データ分析の結果の考察がより広い視点から行うことができると考える。

3.5　音声と検索システム

I-JAS の発話データは書き起こした後，プレインテキストだけでなく検索システムに搭載して，文字列検索および形態論情報を基にした検索が可能な形で公開している。また，発話データについては音声データも併せて公開しているため，書き起こしのテキストの音声・音韻情報を容易に確認することができる。

現在，2016 年 5 月に第一次データ，2017 年 5 月に第二次データを公開し，合計 400 名の学習者と 50 名の日本語母語話者のデータを公開しており，2018 年春に 210 名，2019 年に 215 名を加え，2020 年に 1050 名のデータ公開を予定している（http://lsaj.ninjal.ac.jp/）。

4．言語環境と日本語の習得

4.1　安全な誤用と危険な正用

学習者には，（6）のような文法上，誤用とされる発話や作文が見られ，多くの第二言語習得研究では，これらのデータ分析が行われてきた。しかし，いずれの例も正用はすぐ連想が可能で，発話者の意図は伝わる。一方，（7）の発話例は，文法的な誤用は見られないが，場面によっては聞き手に不快感を与える可能性があり，正用であっても危険な表現となる。

（6）a. ○○さんがお客さまと電話を望むでしたけど，お客さまの外出でメッセジをのこります。（韓国のホテルのボーイの書面による伝言）
　　b. あのときのこと，覚えない，高校生，よく覚えた（中国語話者・女性）
　　c. 花を育つ，野菜を育つ，みんなお母さんがした（英語話者・女性）
（7）a. 教師：では，これから調査を始めます。
　　学生：先生，よろしくね。
　　b. 教師：日本では，学校は 4 月から始まります。
　　学生：へぇ，そうなんだ。
　　c. 教師：最近は海外でも和食の店が増えたそうですね。
　　学生：そうそう。私も日本のラーメン，大好き。

（7）の発話は，I-JAS のデータ収集の調査時に見られた発話例であり，調査者である教師と調査協力者である学生の対話が多く，友人同士なら自然な会話表現でも場面が変わると不自然な会話となってしまう。これらの表現は，海外の大学で学ぶ日本語学習者の発話であるが，全員教室環境で学んでいる。

本研究では以上のような特徴を持つ I-JAS のデータのうち，言語環境の異なる学習者データを比較し，その違いを検討する。具体的には，国内の教室環境と海外の教室環境の日本語学習者を比較し，目標言語圏と非目標言語圏での学習者の習得状況の違いを見る。

さらに，国内の教室環境と国内の自然環境の学習者を比較することで，教室指導の影響を検討する。自然環境の学習者とは，大学や日本語学校などの教育機関に通わずに，日本人の配偶者となったり，就労目的で日本に在住したりしている学習者を指す。しかしながら，彼らの中にも 1 ヵ月に数回，ボランティア教室に通っている学習者もいるので，厳密には，まったく指導が入っていないとは言えない。

4.2　言語環境の違いと習得研究

迫田（2002）は，日本語の習得研究を調べた結果，研究によって，目標言語圏で学ぶ JSL（Japanese as a Second Language）の学習者と自国（非目標言語圏）で学ぶ JFL（Japanese as a Foreign Language）の学習者の違いについて，先行研究をまとめている。その結果をまとめたものが，表 2 である。

また，環境別の習得研究では，Pica（1983）のスペイン人の英語学習者を対象として行った研究がある。Pica は，メキシコの学校で英語を勉強しているグループ 6 名と，アメリカで自然な環境で学校に通わずに習得しているグループ 6 名と，両者をミックスした環境，つまりアメリカの学校で勉強しているグループ 6 名の 3 種類の環境での習得状況を調査した。その結果，表 3 のように，教室指導を受ける学習者は形態素の使用に注目する傾向があるのに対し，自然環境では意味に注意が向けられる傾向があった。この結果から，言語環境の違いが習得に影響を及ぼす可能性が高いことが示された。

表 2　JSL 学習者と JFL 学習者の習得の比較研究

JSL と JFL で違いがあった研究	JSL と JFL で違いがなかった研究
□許（1997）「テイル」では，JSL の正答率が高い。 □田中（1997）文完成テストで JSL は，モダリティの正答率が高い。	■鎌田（1993, 2000）伝達表現の習得では，両者に違いがなかった。 ■稲葉（1991）条件表現の文法性判断調査をした結果，違いがなかった。 ■田中（1997）授受表現や複文の習得には，違いがなかった。

表 3　環境の違いにおける文法形態素習得の特徴（Pica 1983 に基づく）

教室環境	自然環境	ミックス環境
I don't understanding ～のように ing や -ed を過剰に使用する傾向が強い。	two book や many town のように統語的な規則を適用しないで省略する傾向が強い。	低い成績の学習者は自然環境に近く，高い成績の学習者は学校環境に近い傾向を示した。

4.3　教室指導と習得研究

Pica（1983）では，教室で体系的に学ぶ学習者は，文法形態素を意識し，過剰に使用する傾向が見られるという結果を述べている。(8) は，日本の工場で働きはじめて 10 カ月になる自然環境のマレーシア人日本語学習者（男性）の発話例である（S1: マレーシア人の日本語学習者 NS: 日本語母語話者を表す）。

（8）S1：私たちが心配，A ちゃんの言葉が，今，A ちゃん，わからないの，マレーシア語と日本語。
NS：あーそう，バイリンガルだね。
S1：（略）shopping とか super market と，だから（＝なぜなら surrounding が，かん，環境が，日本人いっぱいの，日本語，合わさなきゃいけないの。（＝周囲に日本人が大勢いるので日本語を使って合わせなければいけない。）

森本（1998）は，日本語学校の留学生 3 名とフィリピン人の外国人花嫁 3 名の日本語の習得状況を比較した結果，表 4 のような違いを示している。調査対象の 6 名は，OPI（Oral Proficiency Interview）によって中級レベルと評定されており，口頭能力には大きな差がないことがわかっている。

表 4　教室環境と自然環境における習得状況

教室環境中心の中級レベル学習者	家庭環境中心の中級レベル学習者
平均学習歴　1 年 6 か月	平均滞日歴　約 6 年 8 か月
特徴： ①「です」「ます」の単文が多い。 ②「終助詞」はあまり使用しない。 例　学校へ行きました。	特徴： ①接続詞のある複文が多い。 ②終助詞「よ・ね」を多用する。 例　学校，行ってネ，見たヨ。

ラーセン・フリーマン＆ロング（1995），山岡（1997）は，教室環境と自然環境の習得の違いに関する先行研究をまとめ，指導の効果を（9）にまとめている。

（9）

a. 教室指導の順序は，第二言語の習得順序や発達順序には影響を与えない。
b. 教室指導は，習得の速度を速め，より高いレベルの熟達度を保証する。
c. 教室指導は，自然なスピーキングにつながる効果を持たないが，文法などの正確さを向上させる効果がある。

4.4 「中途終了文（言いさし）」の分析

本稿では，I-JAS の「依頼」のロールプレイの発話を取り上げて，異なる言語環境における習得傾向の違いについて明らかにする。具体的には，ロールプレイにあらわれた（10）のような発話開始時の表現の違いを比較する。

（10）【日本語母語話者】
あのう，すみません，店長さん，ちょっと，<u>お話したいことがあるんですが・・・</u>
【日本語学習者】
店長さん，ちょっと<u>言いたいことがあります</u>。

「依頼」や「断り」に関する日本語の習得研究の先行研究では，①社会的地位の違いにかかわらず，中途終了文を使わず，直接的に断る傾向があり，②外国人の場合，中途終了文は使用が少なく，完結文で話す傾向が高く，③どの研究においても，原因は母語の影響と考える傾向が高い，という結果が出ている（生駒・志村 1993，鮫島 1998，猪崎 2000）。

また，I-JAS のデータの一部を分析した迫田（2015，2016）では，①社会的地位の違いにかかわらず，中途終了文を使わず，直接的に断る傾向があった②外国人の場合，中途終了文の使用が少なく，完結文で話す傾向が高い③どの研究においても原因は母語の影響と考える傾向が高い，という結果を示している。

4.5　問題の所在

本稿では，以下の 2 点を調査課題とし，迫田（2015，2016）にさらに JSL のデータを加えて，再分析を行う。

課題１　目標言語圏内で学習する場合と自国で学習する場合では，中途終了文の使用に違いが見られるか。

課題２　国内の教室で学ぶ場合と自然環境の場合で，中途終了文の使用に違いが見られるか。

4.6　分析対象

分析対象とするロールプレイに用いられるロールカードの内容は（11）のとおりであり，正確に理解させるために学習者の母語または英語で作成されている。

（11）【ロールカード】

あなたは，日本料理店でアルバイトをしています。（中略）
今は，一週間に三日アルバイトをしています。しかし，忙しくなってきたので，一週間に二日に変更したいと思っています。そこで，店長に言って，三日から二日に変えてもらうように頼んでください。

調査は，迫田（2015, 2016）で分析した JFL の日本語学習者（各 20 名），JSL の教室環境と自然環境（各 20 名）に日本語母語話者（15 名）を加え，計 135 名を対象とした。各群のアルファベットは各群の分類記号で，JFL の場合最初の 1 文字は英語表記の母語の頭文字で，残りの 2 文字は，基本的に ISO（国際標準化機構）の国名コードである（例 FFR: French, FRance）。表 5 に，各群と日本語能力テスト（J-CAT）の平均得点を表 5 に示す。JFL の英語話者と中国語話者群は，JSL の教室環境と自然環境の学習者群と能力的にあまり差がないことがわかる。

表５　各群の人数および学習者の日本語能力テスト（J-CAT）の平均得点

	各群（分類記号）人数	得点	各群（分類記号）人数	得点
JFL	仏語話者（FFR）n=20	164.30	西語話者（SES）n=20	162.30
JFL	英語話者（EAU/EUS）n=20	209.75	中国語話者（CCM）n=20	209.30
JSL	教室環境（JJC/JJE）n=20	210.85	自然環境（JJN）n=20	210.10
	日本語母語話者（JJJ）n=15			

分析対象とする発話は，会話の切り出し部分（開始部）と本題の「依頼」を行う発話部分（依頼部）の表現を取り上げる。具体的な「開始部」と「依頼部」の発話例を（12）に示す。そして，開始部と依頼部の発話を，（13）の文に分類し，使用文の割合を示す。さらに，それぞれの学習者間の中途終了文の使用頻度をカイ二乗検定により比較する。

（12）あのー，ご相談があるんですが・・・。　←「開始部」
えーと，ちょっとね，私は今，えと，私，私の今の勉強は，えと，授業が多くなりましたけど，忙しくなります，そして，私は，えと，あのー，アルバイトを週二日に変えていただきたいんですが・・・。←「依頼部」

（13）
中途終了文　⇒ **中途**　例　お話したいことがあるんですが・・・
質問文　　⇒ **質問**　例　今，少しよろしいでしょうか
平叙文　　⇒ **平叙**　例　店長，話があります

4.7　分析結果

「開始部」「依頼部」における使用文の割合を図 2 と図 3 に示す。

図 2 の結果から，①母語話者は中途終了文を多用するのに対し，学習者は平叙文の割合が多い（例　店長，お願いが一つある　中国語話者）②中途終了文の使用割合において，JFL と JSL 学習者に違いがみられる③母語の違いは，中途終了文の使用に大きな影響を与えているとは言い難い，の 3 点が明らかになった。

次に，本題である「依頼」が発話に見られる部分の文をカテゴリーで分けると，図 3 のとおりになる。依頼部の文のグラフの結果から，①環境によって中

途終了文の使用が異なり，国内自然の学習者は中途終了文の使用については，母語話者に最も近い②国内自然以外では，国内教室も含め，学習者は中途終了文ではなく，質問文で依頼を表現する傾向がある（例　すみませんが，一週間に二日に働いてもいいですか　西語話者）ことが明らかになった。

開始部における中途終了文の頻度は，目標言語圏内で学習するJSL環境と自国で学習するJFL環境との間において有意水準5％で差があり（x^2=5.90, df=1, p<.05），JSL環境のほうが有意に多かった。また，依頼部ではJSL自然環境とJSL教室環境との間において有意水準1％で差があり（x^2=6.83，df=1, p<.01），自然環境のほうが有意に多かった。

【開始部】（数字％）

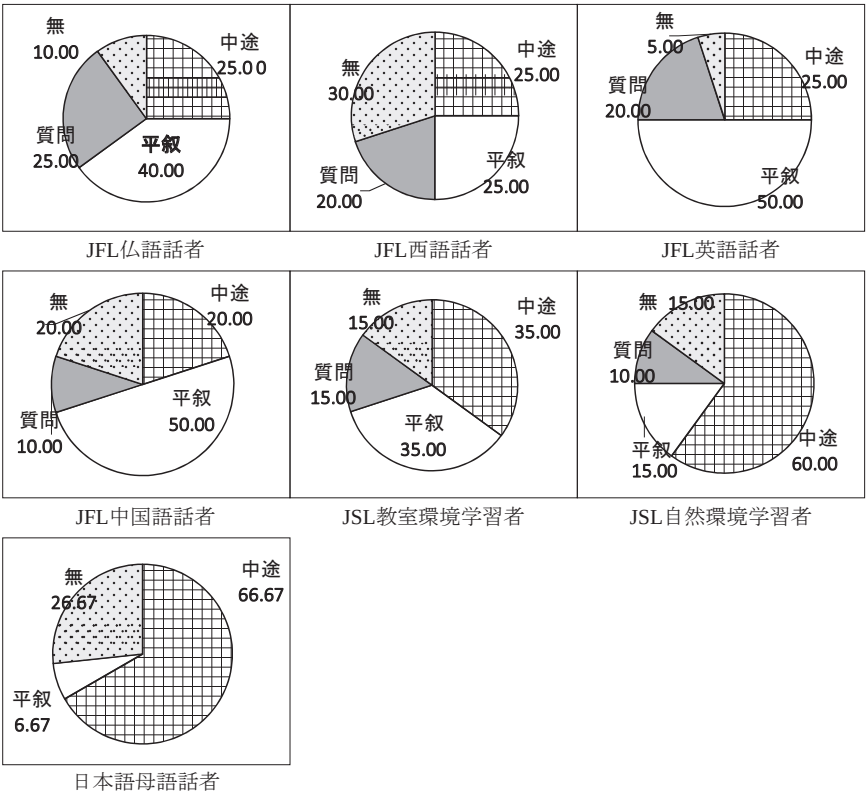


図２「開始部」で使用された文の種類の割合

【依頼部の結果】（数字％）

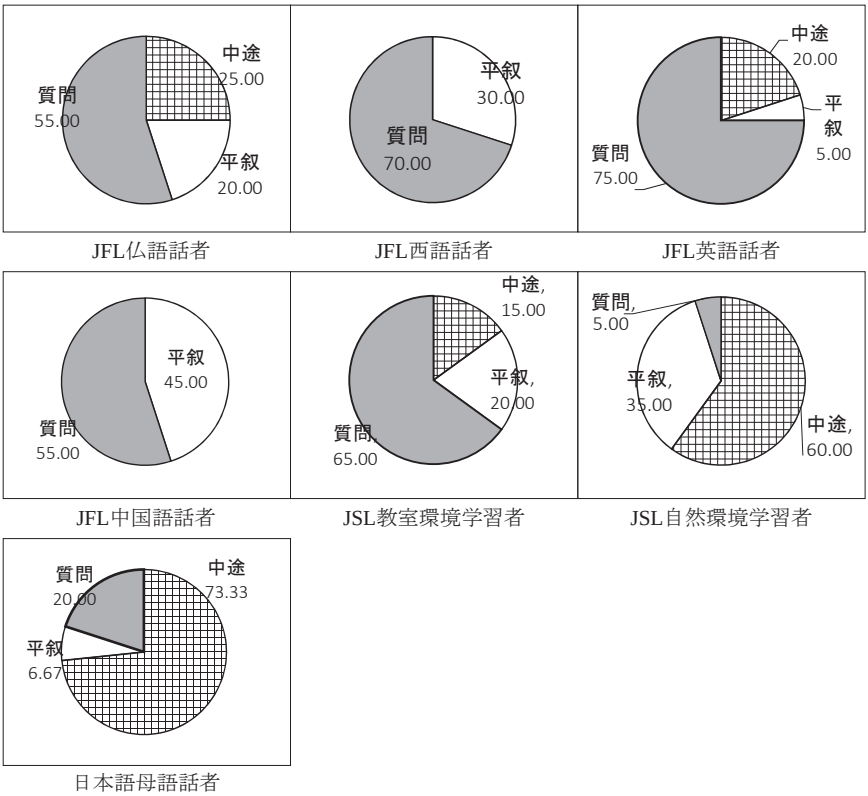


図３　「依頼部」で使用された文の種類の割合

4.8　結論

本分析の調査課題に沿って，明らかになった結果を示し，考察を加える。以下，課題ごとに明らかになったことを示す。

調査課題1：JFL 学習者と JSL 学習者の言語使用の違い

JFL 学習者と JSL 学習者の言語使用においては，開始部では JSL の学習者は中途終了文の使用頻度が JFL 学習者の群よりも有意に多かった（p<.05）。

調査課題2：教室環境学習者と自然環境学習者の言語使用の違い

依頼部においては，JSL 自然環境学習者のほうが JSL 教室環境学習者に比べ，

中途終了文の使用頻度が有意に多く（p<.01），より母語話者に近い傾向が見られた。このことから，中途終了文のような表現は，自然なコミュニケーション場面から習得される可能性が高いことが推測される。

5. おわりに

本稿の目的は，コーパスの重要性と共に日本語学習者コーパスの I-JAS を紹介し，「依頼」の際に母語話者が多用する「中途終了文」を対象として，言語環境の違いによる使用の特徴的な傾向を探ることであった。

自国（JFL）環境と目標言語環境（JSL）において，JFL 学習者と JSL 学習者の中途終了文の使用頻度を比較すると，依頼の開始では，JFL 学習者と JSL 学習者に若干違いが見られ，環境の違いが中途終了文のようなコミュニケーション場面での日本語使用に影響を与えることがわかった。

さらに，目標言語環境（JSL）内の教室環境学習者と自然環境学習者の使用の違いを検討した。グラフを見る限り，中途終了文に関しては，JSL 自然環境学習者の言語使用が日本語母語話者に近く，JSL 教室環境の学習者は海外（JFL の日本語学習者の傾向に近いことがわかった。このことから，中途終了文のような表現は，教科書で指導するだけでなく，自然なコミュニケーション場面で習得させる必要があることがわかった。中途終了文は，日本語指導の項目の対象となっていない場合が多く，今後の日本語指導を検討する上で「正確かどうか」だけでなく「適切かどうか」の観点が重要となる。

今回は，中途終了文の使用に注目して，データの分析を行ったが，他の項目，たとえば教室指導で教えられる「授受表現」「受身」などを対象とした場合は，違う結果が出てくる可能性がある。今後は，さらにデータを増やし，対象項目を検討し，言語環境の違いが日本語習得に及ぼす影響について研究を進めたい。

本研究は，科学研究費（課題番号 16H01934）「海外連携による日本語学習者コーパスの構築と言語習得および教育への応用」の助成を受けて行った。

参考文献

生駒知子・志村明彦（1992）「英語から日本語へのプラグマティック・トランスファー―「断り」という発話行為について―」『日本語教育』79，pp.41–49.
市川保子（1997）『日本語誤用例小辞典』凡人社.

稲葉みどり（1991）「日本語条件文の意味領域と中間言語構造」『日本語教育』75 号，pp.87–99.
猪崎保子（2000）「「依頼」会話にみられる『優先体系』の文化的相違と期待のずれ」『日本語教育』104，pp.79–87.
鎌田修（1993）「日本語の中間談話文法の一側面」『日本語・日本文化研究』創刊号 pp.14–28，京都外国語大学留学生別科.
鎌田修（2000）『日本語の引用』ひつじ書房.
許夏珮（1997）「中上級台湾日本語学習者による「テイル」の習得に関する横断研究」『日本語教育』95 号，pp.37–48.
迫田久美子（1998）『中間言語研究―日本語学習者における指示詞コ・ソ・アの習得研究』渓水社.
迫田久美子（2001）「学習者の誤用を産み出す言語処理のストラテジー（1）―場所を表す「に」と「で」の場合―」『広島大学日本語教育研究』11，pp.17–22，広島大学大学院教育学研究科日本語教育学講座.
迫田久美子（2002）『日本語教育に生かす第二言語習得研究』アルク.
迫田久美子（2015）「学習者のロールプレイに見られる話し手の依頼表現：日本語学習者コーパスにおける対話 ―ロールプレイ，メール，エッセイの分析をとおして―」田中真理・野田尚史とパネル発表『ヨーロッパ日本語教育』20，pp.102–107．ヨーロッパ日本語教師会.
迫田久美子（2016）「学習者の発話データに基づく日本語の習得：学習者のロールプレイに見られる話し手の依頼表現（2）―レベル差の観点から―」ICJLE2016，パネル発表.
file:///C:/Users/sakodak/AppData/Local/Microsoft/Windows/INetCache/IE/JQEEVJHV/ICJLE2016_JP_Panel- 提出版）8_11.pdf
迫田久美子・小西円・佐々木藍子・須賀和香子・細井陽子（2016）「多言語母語の日本語学習者横断コーパス」『国立国語研究所 プロジェクトレビュー』6，3，pp.93–110.
鮫島重喜（1998）「コミュニケーションタスクにおける日本語学習者の定型表現・文末表現の習得過程―中国語話者の「依頼」「断り」「謝罪」―」『日本語教育』98，pp.73–84.
田中真理（1997）「視点・ボイス・複文の習得要因」『日本語教育』92 号，pp.107–118.
寺村秀夫（1990）『外国人学習者の日本語誤用例集（資料集）』（1985–1989 年度文部省科学研究費特別推進研究「日本語の普遍性と個別性に関する理論的及び実証的研究」の分担研究）
野田尚史・迫田久美子・渋谷勝己・小林典子（2001）『日本語学習者の文法習得』大修館.
福間康子（1997）「作文からみた初級学習者の格助詞「に」の誤用」『九州大学留学生センター紀要』第 8 号，pp.61–74．九州大学留学生センター.
森本智子（1998）「異なった学習環境における日本語習得の違いに関する研究―教室

環境と自然環境の学習者を対象として―」平成 9 年度広島大学教育学部日本語教育学科卒業論文
山岡俊比古（1997）『第 2 言語習得研究＜新装改訂版＞』桐原ユニラーセン－フリーマン，D. &ロング，M.（1995）『第 2 言語習得への招待』（牧野高吉（訳））鷹書房弓プレス［Larsen-Freeman, D. and Long, M. (1991) An Introduction to Second Language Acquisition Research. London: Longman］
Bley-Vroman, R. (1983) The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning 33*.1–17.
Pica, T. (1983) Adult acquisition of English as a second language under different conditions of exposure. *Language Learning 33*: 465–497.

（迫田久美子　広島大学　Email: sakodak@hiroshima-u.ac.jp）
（細井　陽子　国立国語研究所　Email: y.hosoi@ninjal.ac.jp）

「シンポジウム」
話し言葉コーパスの構築と利用
JECPRESE: JSL と EFL ユーザーのために

野口ジュディー

1．はじめに

JECPRESE, the Japanese-English Corpus of Presentations in Science and Engineering（http://www.jecprese.sci.waseda.ac.jp/）は留学生のための専門日本語教育（JSL, Japanese as a second language）を支援する研究発表コーパスとしてスタートした。日本の大学院生の日本語プレゼンテーションに加えて，米国の大学生や国際学会の英語プレゼンテーションも収められており，EFL（English as a Foreign Language）の学生にも利用可能なコーパスとなった。理工系のプレゼンテーション（発表）の特徴をわかりやすくするために, 各発表について, ESP（English for Specific Purposes）の手法であるジャンル分析に依拠したセクションやステップで検索することができるように設計してある。単語や表現の検索もできるようになっている。

2．プレゼンテーションコーパスの必要性

科学技術分野では，研究の成果を専門分野の研究者と共有することが必須である（Gross 1990, Myers 1990, Canagarajah 2002）。研究成果を共有する方法としては，専門誌への論文投稿や学会での口頭発表，ポスター発表などがある。論文の書き方についての教科書は数多く出版されており，また，論文テクストによるコーパスを構築すると，執筆時に参考にすることができる。しかし，口頭発表やポスター発表の資料は入手するのが難しい上に，コーパス化するためには音声を書き起こすことが必要になる。この問題を解決するために，研究発表の日本語と英語を集めたコーパスを構築した。

最初に，日本語による研究発表をコーパス化することから着手した。日本の大学には，科学技術を学ぶために海外から訪れる留学生が多いため，日本語での研究発表コーパスがあれば，各自の研究発表のみならず，他の学生による日

本語での発表を聞く際の助けにもなるであろうと考えられた。これは，後述する ESP（English for specific purposes）のジャンルの概念に基づくものである。

3. ESP（English for specific purposes）のジャンル分析

グローバル規模で科学技術研究者が高度なレベルでの情報交換を行うには，科学の内容を正確に伝えるための英語力が必要となる。しかし，科学技術を専攻する学生や研究者の全員が高い言語運用能力を有するとは限らない。良い研究を行っても，研究内容を上手に発表することができないと研究に対する理解を得ることは難しいであろう。そのような状況で，学生や研究者の助けになるアプローチが専門英語研究（ESP）から生まれたジャンル分析の方法である（Swales 1990）。

ESP では，英語運用能力全般を引き上げることを目標としない。そのことより，特定の文書や話し言葉の特徴を明らかにして，その使い方の向上を目指す。例えば，パーティーで流暢に会話を行うことより，良い研究論文を書くことを優先する。それは，文書の専門性が高ければ高いほど，より定型的なパターンが認められ，学びやすいためである。

このパターン化を簡単な例で検証してみよう。学会での口頭発表を行いたい場合，まず，発表内容を把握できるような要旨を学会に提出する。この内容が学会の趣旨と合致し，良い研究であると判断されると，口頭発表を行うことが許可される。最近，ハワイで開催された国際学会での口頭発表の要旨について，ジャンル分析（パターン分析）をしてみよう。

冒頭の括弧内には，それぞれのセクションの目的が書いてある。文中に下線を施した部分は，読み手にヒントを与える表現である。これらの表現を手がかりにすると，書き手がそのセクションで何を伝えようとしているのかが理解しやすくなる。このように，文章の目的をわかりやすくすることで文書が読みやすくなる（Tojo et al. 2014）。

> 【Purpose of presentation】 **This paper presents** corpus linguistic **findings that show** how differences between Japanese and English influence the way in which science concepts are transmitted in science education today.【Background】 **We first review** how Western science concepts entered Japan via translation from Dutch during the Edo Period (1603 to 1868) and then the impact on science education after World War II of American science textbooks being translated for

classroom use. **Today,** with increasing globalization, the Japanese Ministry of Education, Culture, Sports, Science and Technology has designated Super Science High Schools and encouraged the delivery of lectures in English at universities. 【Research aim & methods】 **To find ways to help** instructors prepare their lectures in English and to support students who need to listen to them, since 2010, **we have been working with** corpora of university engineering lectures presented in English. **However**, due to the language distance between English and Japanese, communication differences between Western and Japanese cultures, and prevalent teaching traditions in Japan, **we became aware of the need to** compare the English lectures with those delivered in Japanese. From 2016, being interested in how scientific knowledge was being imparted to the students, we began adding lectures delivered in Japanese to our corpus.【Results】 **Comparison of** these Japanese lectures with the English ones **has revealed differences** in the lecturers' perceptions of how science should be taught.【Discussion】 **We discuss how** language and cultural differences can influence the way concepts are transmitted in science education. (Noguchi, J., Kunioshi, N., & Tojo, K. Language and cultural influences on "doing science in Japanese", The IAFOR International Conference on Education – Hawaii, January 4 to 6, 2018, Honolulu)

このように，明確な目的がある文書に用いられる表現には，繰り返し使用されるうちにより効率よく内容が伝わるようにパターン化されたものが多い（Dressen-Hammouda 2008）。書いたり話したりして発信するときも，読んだり聞いたりして受信するときも，専門分野の書き言葉や話し言葉が有するパターンと高頻度に使用されるヒント表現を知っていると，受発信することが容易になる。パターンを知るための方法のひとつに，コーパスを利用して，頻回に使用される表現を検索することがある。ESP の概念に基づく文書の捉え方を理解することができると，必要な専門文書をより平易に学ぶことができるようになる。

4．アカデミックプレゼンテーションコーパスの構築

4.1　JECPRESE の日本語プレゼンテーションコーパス

日本の工学系大学院で研究を行う留学生を支援する目的で，日本人学生による日本語の口頭発表を 2006 から 2008 年にかけて録音し，書き起こしを行った。

データ収集は，各大学院の各専攻に許可を得た上で，発表者個人からも許可書を得て行った。収録を許可されたコーパスにおけるデータの内訳は，物質科学が 34，分子化学が 31，知能機能創成工学が 34，電気工学が 13，機械工学が 61 であり，日本語の話し言葉（理系プレゼンテーション）としては約 120 万語であった（Kunioshi et al. 2012）。

4.2 JECPRESE の英語プレゼンテーションコーパス

英語を第二言語として使用する学生のためには，英語でのプレゼンテーションのコーパスが役立つであろうと考えられたため，2 種類（2 つのジャンル）の英語の口頭発表をコーパスに加えた。1 つ目は，化学分野の国際学会おける，さまざまな国のベテランの研究者による口頭発表であり，データ数は 45，英語の話し言葉（化学系プレゼンテーション）としては約 29 万語であった。2 つ目は，米国の大学の，工学部の学生によるグループ発表であり，データ収集に際しては，約 1 年かけて大学の倫理委員会に収録の許可を得た上で，学生個人の承諾書も取得した。収録の対象となった学生数は，グループごとに 3 〜 4 名，合計 54 名であり，データの内訳は，電気工学が 2，土木・環境工学が 5，機械工学が 6，コンピュータ科学が 2 であり，英語の話し言葉（理系プレゼンテーション）としては，約 44 万語であった。

5. JECPRESE の検索システム（インタフェース）

収集した話し言葉日英（理系プレゼンテーション）コーパスは，利便性を考慮して Fig. 1 のような検索システム（インタフェース）からアクセスできるように整備した。検索窓に言葉を入力すると，コンコーダンスラインを得ることができる。しかし，英語や日本語の発表に不慣れな学生にとっては，入力する言葉を思い付くこと自体が難しいと考えられたため，セクション（Sections）とステップ（Steps）という選択肢を設けた。セクションとステップの選択肢についての詳細は，Table 1 を参照されたい。例えば，セクションの *S* （Start）のチェックボックスにチェックを入れると，プレゼンテーションの開始時に発話された言葉が表示される。日本語での口頭発表の開始時には，Table 2 に示したような典型的な表現が繰り返し使用されていることが明らかになった。また，プレゼンテーションの終了時には，「以上です。」という表現が 18 件，「発表を終わらせていただきます。」という表現が 14 件認められた。

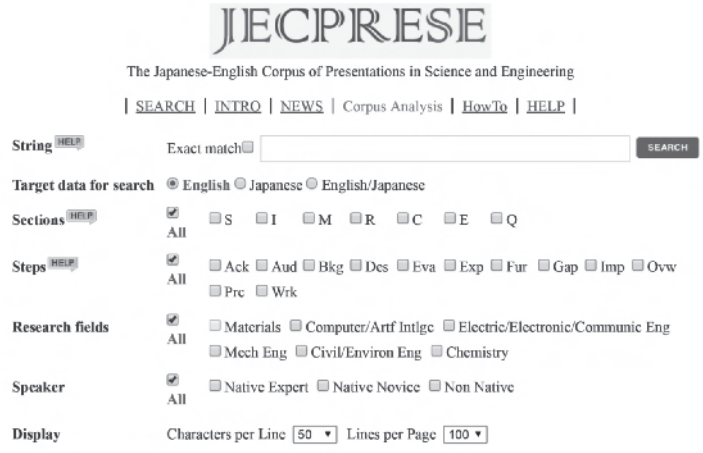


Fig. 1　JECPRESE の検索ページ

Table 1　List moves found in English and Japanese presentations

Tag	Section	Tag	Description
S	Start	Ack	Acknowledgments
I	Introduction	Aud	Audience orientation
M	Materials and methods	Bkg	Background
R	Results and discussion	Des	Description
C	Conclusion	Eva	Evaluation
E	Ending	Exp	Explanation
Q	Question and answer	Fur	Further research
		Gap	Gap
		Imp	Implication
		Ovw	Overview
		Prc	Procedures
		Wrk	Present work

Table 2　日本語での口頭発表の開始時における表現

高頻度使用表現	回数
TITLE と題しまして SURNAME 研究室の SURNAME が発表させていただきます.	16
それでは，発表させていただきます.	14
SURNAME 研の SURNAME です.	5
SURNAME 研究室の SURNAME です.	4

JECPRESE の検索システムでは，口頭発表の各セクション（Sections）におけるステップ（Steps）での詳しい説明の表現を検索することもできる。例えば，英語での口頭発表の導入部で先行研究の隙間（Gap）をどのように説明しているかを知りたい場合，検索対象（Target data for search）の項目で英語のプレゼンテーション（English）をクリックし，セクション（Sections）の *I*（Introduction）とステップ（Steps）の *Gap*（研究の意義を説明するステップ）にチェックを入れると，以下のような例が表示される。学生は，表示された表現を参考に各自のプレゼンテーションを準備することができる。

And **we were also asked to design** a sustainable alternative for the reuse of their grey water.

For our project we were to design an onsite waste water facility to accommodate a new residential

So for our problem statement, we are to construct a machine that can replace the current system for

So they are looking for more durable system to make it easier for the river guides.

The problem with this is that you have two materials that are difficult to separate, as you know,

There are two types of challenges.

What we need, however, is a deeper understanding of complex structures.

なお，日本語の口頭発表では，Gap を述べる際に「しかし，」と「そのため，」が頻用されることが判明した。

6．姉妹サイト OnCAL の紹介

JECPRESE の姉妹サイトとして，JECPRESE と類似した検索システム（インタフェース）を有する OnCAL（The Online Corpus of Academic Lectures）（http://www.oncal.sci.waseda.ac.jp/）がある。OnCAL コーパスのデータ数は，英語での工学系講義［MIT OCW（Massachusetts Institute of Technology, Opencourseware）と SEE（Stanford Engineering Everywhere）］が 430 で，英語の話し言葉（工学系講義）としては約 350 万語（395 講義時間）である（Kunioshi et al., 2015）。収録

データは，OnCAL の検索システムを用いて検索することができる。Fig. 2 に OnCAL 検索システムのホームページのイメージを示した。この検索システムでは, JECPRESE の検索システムとは異なるアプローチで検索を行うことができる。

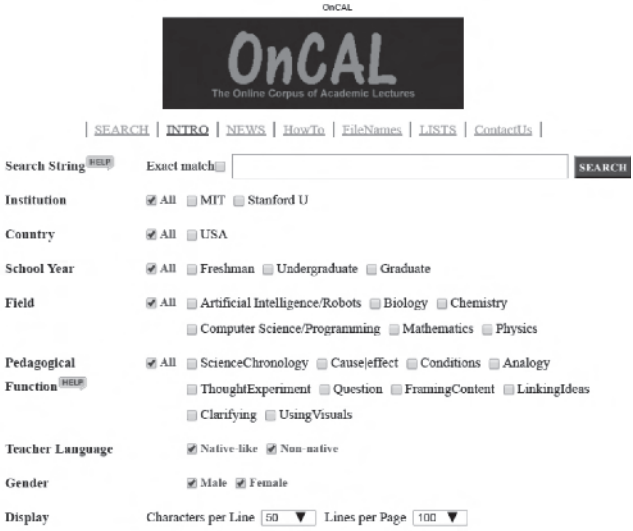


Fig. 2　OnCAL の検索ページ

OnCAL は，英語での工学系講義の流れと表現を明らかにして，日本の大学で英語による理工系の講義を支援することを目的として構築した。OnCAL の検索システムで検索できる項目の中に，Pedagogical Function（教育的機能）がある。例えば，Pedagogical Function の選択肢から，*Thought Experiment* を選択して検索すると，*let's suppose that* や *let's imagine that* などが頻出することがわかる。また，同じく選択肢から，*Cause / effect* を選択して検索すると，*as a consequence of that* や *as a result of this, can lead to, causes it to* などが提示される。このような検索を行うと，英語で講義を行うのに必要な表現を知ることができるだけではなく，英語での講義を聞くのにも役立つであろうと考えられる。

7．終わりに

話し言葉コーパスとして，日英（理系プレゼンテーション）コーパス JECPRESE とその検索システムを紹介し，姉妹サイトとして英語（工学系講義）

コーパス OnCAL についても簡単に紹介した。是非 JECPRESE と OnCAL の検索サイトにアクセスして，試していただきたい。

謝　辞

JECPRESE は，科学研究費助成事業（科学研究費補助金）基盤研究（C）21520601「理工系口頭発表コーパスに基づいた専門日本語・英語の教育法の開発」の支援を受けた。OnCAL は，科学研究費助成事業（科学研究費補助金）基盤研究（B）24300273「英語を介した理工系高等教育の向上を支援するシステムの開発」の支援を受けた。構築に携わった研究員は，国吉ニルソン，東條加寿子，林洋子，野口ジュディーであった。

参考文献

Canagarajah, S. (2002) Multilingual writers and the academic community: towards a critical relationship. *Journal of English for Academic Purposes*, 1 (2002) 29–44.

Dressen-Hammouda, D. (2008) From novice to disciplinary expert: Disciplinary identity and genre mastery. *English for Specific Purposes*, 27, 233–252.

Gross, A. (1990) *The Rhetoric of Science.* Cambridge, MA: Harvard University Press.

Kunioshi, N., Noguchi, J., Hayashi, H. & Tojo, K. (2012) An online support site for preparation of oral presentations in science and engineering, *European Journal of Engineering Education*, 2012, 1–9.

Kunioshi, N., Noguchi, J., Tojo, K. & Hayashi, H. (2015) Supporting English-medium pedagogy through an online corpus of science and engineering lectures, *European Journal of Engineering Education*. http://dx.doi.org/10.1080/03043797.2015.1056104

Myers, G. (1990) *Writing biology: Texts in the social construction of scientific knowledge. Madison*, Wisconsin: The University of Wisconsin Press.

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

Tojo, K., Hayashi, H., & Noguchi, J. (2014) Linguistic dimensions of hint expressions in science and engineering research presentations. *JACET Selected Papers* Vol. 1(2014), 131–163.

（神戸学院大学名誉教授　Email: jnoguchi@gc.kobegakuin.ac.jp）

「シンポジウム」
TED Corpus Search Engine: TED Talks を研究と教育に活用するためのプラットフォーム

長谷部陽一郎

1．はじめに

本稿では TED（https://ted.com）で公開されている英語プレゼンテーションのトランスクリプト，翻訳テキスト，動画データをコーパスとして用いるための検索エンジンである TED Corpus Search Engine（TCSE）の概要とその可能性について論じる。まず 2 節では TCSE の主な機能を概観する。TED Talks という素材に特化した検索システムである TCSE には，そのことを生かした，いくつかの特徴的な機能が実装されている。次に 3 節では TCSE に含まれるデータの様々な統計値を示しながら，TED Talks の言語コーパスとしての特徴を素描する。また，TED Talks の言語がどれだけ「話し言葉」としての特徴を有しているか，あるいはそうでないのかという問題について，予備的な考察を示す。続く 4 節では TCSE の理論的な背景について述べる。ここでは TCSE が認知言語学の言語観に基づいて設計されており，その点において他のコーパス・システムと大きく異なっていることを示す。最後の 5 節では全体のまとめを行う。

2．TCSE の主な機能

TCSE（https://yohasebe.com/tcse）の最も重要な機能は TED コーパスを自在に検索し，その結果を様々な形式で表示することである。本節では，1）英語トランスクリプトの検索，2）多言語による対訳の表示・検索，3）検索対象となる言語単位と表示のバリエーションの 3 つに焦点を当て，システムの概要を示す。なお，TCSE はインターネットに接続された機器（PC，スマートフォン，タブレット）上のウェブ・ブラウザから利用可能である。個別の機能に関する詳しい解説は，サイト上で閲覧できるチュートリアル文書などを参照されたい。

2.1 英語トランスクリプトの検索

図 1 は TCSE のメイン検索パネルである。テキストボックスに英語表現を入力して SEARCH ボタンをクリックすると, 図2のように結果が一覧表示される。

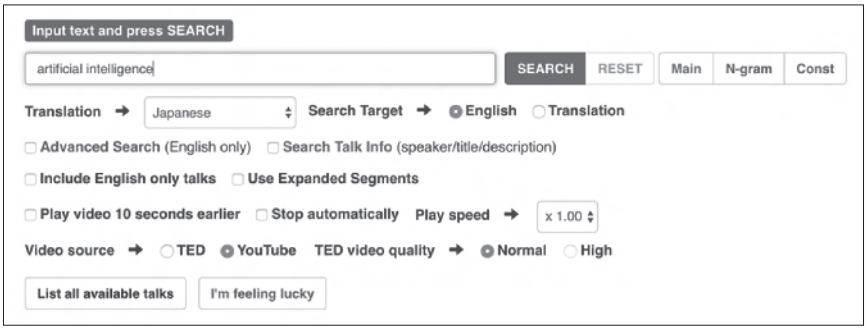


図 1　TCSE の検索パネル

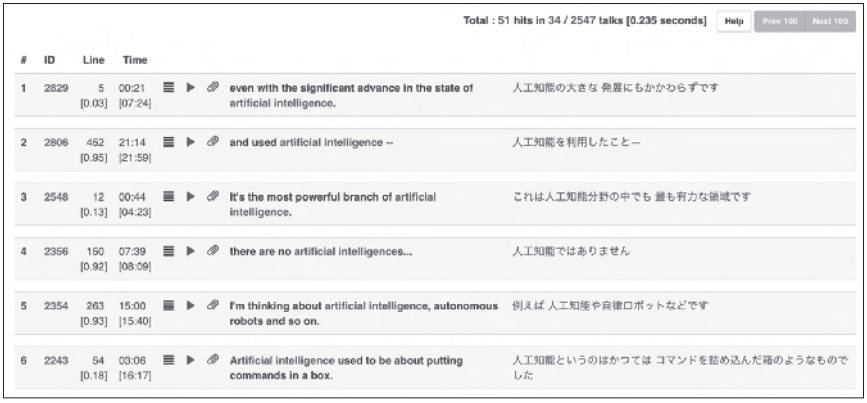


図 2　TCSE の検索結果の例

英語トランスクリプトの検索にあたっては，入力文字列を単純にマッチさせる通常検索に加えて, Advanced Search モードが存在する。Advanced Search モードでは，語の表層形，レンマ，品詞，あるいはこれらを組み合わせた文字列を用いた複合的な検索が可能である。

2.2 多言語による対訳の表示と検索

TED Talks のトランスクリプトは様々な言語への翻訳が進められており，TCSE ではそのうち，「1,000 件以上のトークが翻訳されている」という基準のもとに，日本語を含む 28 言語による対訳データを収録している。表示する対訳言語は検索パネル上で選択可能であるが，指定がない場合はウェブブラウザの設定に応じた言語が選択される仕様になっており，日本語環境では通常，図 2 のように日本語の対訳が英語のトランスクリプトと共に表示される（当該のトークが翻訳済みである場合）。また，対訳テキストを検索することも可能である。対訳言語を選択した上で，Search Target を Translation に設定すれば，テキストボックスに入力された内容をその言語の文字列と見なし，一致する表現をリストアップする。図 3 は日本語による検索結果の例である。

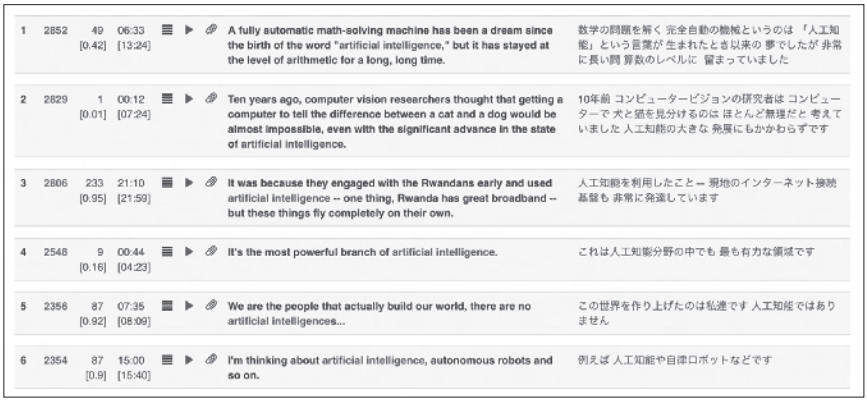
図３　日本語による検索結果の例

2.3 TCSE における検索対象の単位と表示のバリエーション

TED で公開されているトランスクリプトのデータは，1 画面に表示する字幕の幅を基準に構成されており，TCSE でも基本的にこれを採用している。つまり，TCSE において検索結果としてデフォルトで表示されるのは，入力テキストと一致する語句を含んだ 1 画面分の字幕文字列である。TCSE ではこの単位をセグメント（segment）と呼んでいる。

しかしながら，セグメントは往々にしてセンテンスに満たない単位であり，ユーザーとしては表示される個々の結果をより詳しい形で見たい場合があり得る。また，長めの文字列を入力したときなど，セグメント単位では期待された

表現が結果に含まれない場合がある。そこで TCSE では，拡張セグメント（expanded segment）という単位での検索・表示のオプションを設けている。拡張セグメントとは，セグメントがセンテンスの断片である場合，隣接するセグメントと連結し，センテンス全体が同じ要素内に収まるよう調整した単位である。図 4 は拡張セグメント・モードでの検索結果を示したものである。



1	2852	49 [0.42]	06:33 [13:24]	A fully automatic math-solving machine has been a dream since the birth of the word "artificial intelligence," but it has stayed at the level of arithmetic for a long, long time.	数学の問題を解く 完全自動の機械というのは 「人工知能」という言葉が 生まれたとき以来の 夢でしたが 非常に長い間 算数のレベルに 留まっていました
2	2829	1 [0.01]	00:12 [07:24]	Ten years ago, computer vision researchers thought that getting a computer to tell the difference between a cat and a dog would be almost impossible, even with the significant advance in the state of artificial intelligence.	10年前 コンピュータービジョンの研究者は コンピューターで 犬と猫を見分けるのは ほとんど無理だと 考えていました 人工知能の大きな 発展にもかかわらずです
3	2806	233 [0.95]	21:10 [21:59]	It was because they engaged with the Rwandans early and used artificial intelligence -- one thing, Rwanda has great broadband -- but these things fly completely on their own.	人工知能を利用したこと― 現地のインターネット接続基盤も 非常に発達しています
4	2548	9 [0.16]	00:44 [04:23]	It's the most powerful branch of artificial intelligence.	これは人工知能分野の中でも 最も有力な領域です
5	2356	87 [0.92]	07:35 [08:09]	We are the people that actually build our world, there are no artificial intelligences...	この世界を作り上げたのは私達です 人工知能ではありません
6	2354	87 [0.9]	15:00 [15:40]	I'm thinking about artificial intelligence, autonomous robots and so on.	例えば 人工知能や自律ロボットなどです

図４　拡張セグメントを用いた結果表示

セグメントを用いるにせよ，拡張セグメントを用いるにせよ，個々の具体的な表現を詳しく確認する必要がある際には，トークの全文テキストを対訳と共に表示することが可能である。また，検索結果のそれぞれについて，発話箇所の映像をピンポイントで再生することができる。これは従来の多くの言語コーパスには見られない重要な特徴である。

以上，簡単にではあるが，コーパス検索システムとしての TCSE の機能を概観してきた。次節では，TCSE に収録されている TED Talks のデータがコーパスとしてどのような性質を持っているのかについて考える。

３．コーパスとしての TED Talks

TCSE は TED Talks のデータをコーパスとして用いるためのオンライン・システムであり，2014 年の 11 月に最初のバージョンが公開された（Hasebe, 2015）。その後，数多くの機能を実装すると共に，新たなデータの追加を定期的に行ってきた。本節ではこうした「TED コーパス」の特徴を概観する。

3.1 TED Talks のデータについて

TED では現在 2,600 以上の英語プレゼンテーションが公開されており，そのデータは Creative Commons ライセンス（CC BY-NC-ND）のもとで利用可能になっている。CC BY-NC-ND とは，再出典を明記し（BY），非商用（NC）で，内容の改変を行わないという条件のもとにデータの再配布（ND）を認めるタイプのライセンスであり, TCSE もこれに準拠した形でデータを利用している。

TED が主催するカンファレンスにはいくつかの種類があるが，公式のイベントの他に，TEDx と呼ばれる非公式のカンファレンスが存在する。TEDx は TED からライセンスを受けた各地のコミュニティーが独自に開催するもので，厳密には TED カンファレンスと区別される。しかし，その一部は TED の公式サイトでデータが公開されており，上記の「2,600 以上」というプレゼンテーション数にも含まれている。TCSE では，TED により公開されている全データのうち 2,547 件のデータを収録している（2018 年 2 月現在）。収録対象となっていないプレゼンテーションがあるのは，英語以外の言語で話されているものや，トランスクリプトの形式が TCSE のシステムに合致しないものを除外しているためである。

なお，TED ではプレゼンテーションの動画に加えて，発話内容のトランスクリプトを公開しているが，トランスクリプトの多言語への翻訳はボランディア・メンバーの作業によるものである。組織としての TED は翻訳の作業自体には携わらないが，コミュニティーとして翻訳・校閲作業が円滑に進められる仕組みを構築している。登録済みボランティアメンバーによる翻訳テキストは別の登録メンバーによる校閲を受け，これをクリアした翻訳のみが公開されることになっている[1]。

3.2 基本統計値

表 1 に TCSE が収録している TED Talks データの基本統計値を示す。

表 1　TCSE に収録された TED Talks の基本統計値

項　目	総　数
トーク	2,547
セグメント	691,788
拡張セグメント	308,720
エレメント（延べ）	6,166,375
エレメント（異なり）	83,909
語（延べ）	5,306,719
語（異なり）	83,891

それぞれの項目について簡単に解説したい。トーク（talk）とは個々のプレゼンテーションを指す。セグメント（segment）とは，TCSE における基本的な言語単位であり，トランスクリプトにおける「1 画面分の字幕文字列」に一致する。セグメントは「センテンス未満」の単位であることが多いが，そのようなセグメントを隣接するセグメントと結合し，センテンスを 1 つ以上含むセグメントに拡張したものが拡張セグメント（expanded segment）である。エレメント（element）は単語と類似した単位であるが，トランスクリプト内に含まれるメタ表示（Applause や Laughter など）や，句読点などの記号類も含んでいる。したがって，これらはその下の「語（延べ）」や「語（異なり）」よりも大きな値になっている。

3.3 継続時間および発話速度

コーパスとしての TED の 1 つの特徴は，いずれもおよそ 10〜20 分程度でそれぞれが完結した内容のトークとなっていることである。また，スピーカーの多様さには目を見張るものがある。1 人のスピーカーが複数のトークを担当することもあるが，TCSE に収録された 2,547 のトークは，実に 2,166 組の異なるスピーカーによるものである。しかし，そのように多様なスピーカー達によるトークであっても，継続時間や発話速度は，全体を通じてある程度の統制がみられる。

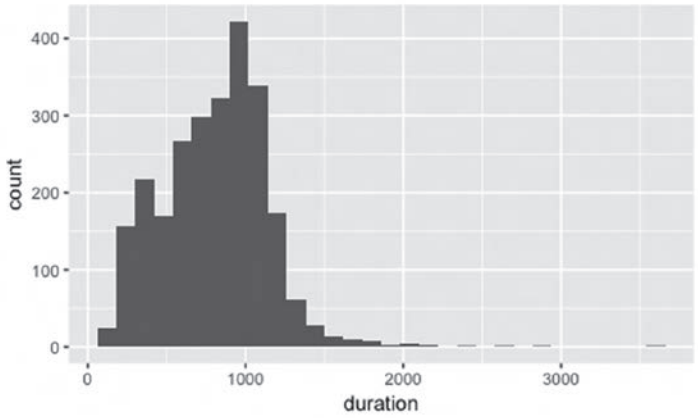


図 5　TCSE に収録されたトークの継続時間（秒）分布

図 5 は継続時間（秒）の分布を示したものである。いくらかの外れ値がみられるものの，それらを除けば 1,000 秒に満たない位置にピークを持つ分布を成していることが分かる。全トークの平均継続時間は 802.12 秒（SD = 330.57），すなわち約 13 分である。

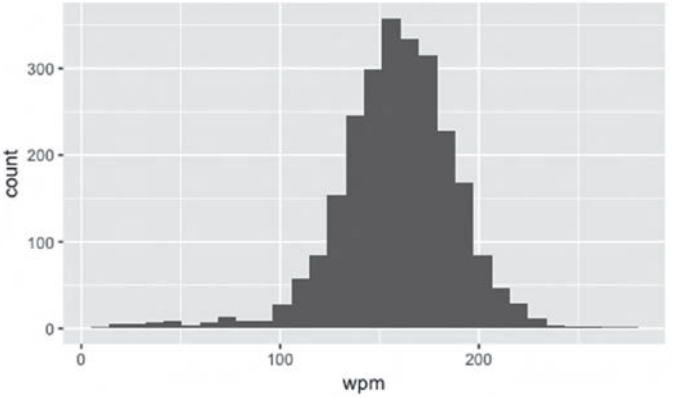


図６　TCSE に収録されたトークの平均発話速度分布

次に図 6 に発話速度（wpm）の分布を示す。平均発話速度は 158.20 wpm（SD = 30.47）である。なお，TED には舞踏や音楽を含むトークも多く含まれており，一部のトークについては正確な発話速度を計測することが難しい。図 6 で 100 wpm 以下に位置するデータは概ねそのようなものである。しかし，いずれにしても全体的には 160 wpm あたりをピークとした分布を示す。Carpenter and Just（1977）は英語の標準的な発話速度を 160 wpm としており，TED の平均発話速度はこれに一致する。

3.4　難易度

TCSE には言語教育に資する様々な機能が実装されているが，その 1 つはトークの相対的な難易度指標の表示である。この機能は Flesch-Kincaid Readability Ease（以下，Flesch-Kincaid 値とする）を計算することによって実現されている。Flesch-Kincaid 値は書かれた文章の「読みやすさ」を評価するための指標であるが，TCSE では多数のトークを教育的な目的で選り分ける際の参考になる要素の 1 つとして本指標を採用している。Flesch-Kincaid 値の計算式は次の通りである。

206.835−1.015（総語数 / 総文数）−84.6（総音節数 / 総語数）

図 7 は TCSE に収録された TED Talks のデータにおける Flesch-Kincaid 値の分布を示したもので，平均値は 57.27（SD = 10.16）である。なお，この平均値を伝統的な Flesch-Kincaid 値の評価の目安に照らすと，Fairly Difficult（高校レベル）と評価される（Gray, 2012）。

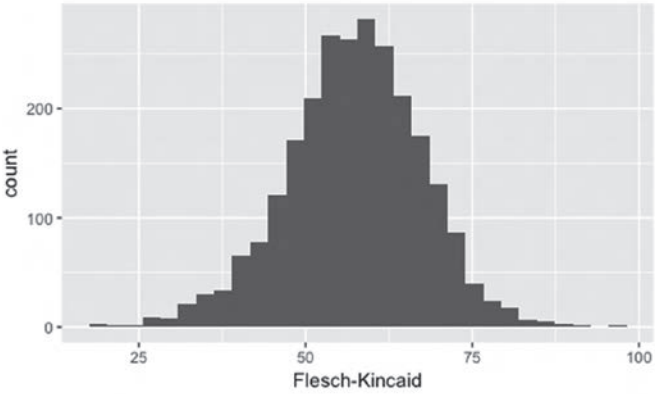


図 7　TCSE に収録されたトークの Flesch-Kincaid 値分布

3.5　TED コーパスの言語的特徴

ここでは，特殊コーパス（specialized corpus）としての TED コーパスの言語的特徴について予備的な考察を行う。一般的に言語コーパスは書き言葉（written language）から構成されたものと話し言葉（spoken language）で構成されたものとに大別される。TED Talks は英語の話し言葉によるプレゼンテーションであり，第一義的には話し言葉コーパスの一種として分類されるべきである。しかし，日常会話の言語を話し言葉の典型とするなら，聴衆の前で準備された内容を披露するプレゼンテーションの言語はそのような典型からはいくらか離れたバリエーションと考えるのが妥当であろう。では，TED Talks の言語は典型的な話し言葉から「どれだけ」また「どのように」離れているのだろうか。現在のところ，この問題に即答することは難しい。ここでは予備的な見通しを得るために，大規模コーパスとの定量的な比較調査の結果に基づいた考察を示す。

Corpus of Contemporary American English（COCA）は異なる使用域の現代アメリカ英語から成る大規模コーパスである（https://corpus.byu.edu/coca）。COCA では，話し言葉（spoken），フィクション（fiction），一般雑誌（popular

magazines），新聞（newspapers），学術雑誌（academic journals）の 5 つの使用域のデータがほぼ均等に（各 20 パーセント）含まれるよう調整されている（Davies, 2010）。そこで，これら 5 つの使用域のサブコーパスから抽出した COCA の語彙データ（レンマ＋頻度）と，TCSE から抽出した TED Talks の語彙データ（レンマ＋頻度）とで順位相関係数（Kendall's *tau*）を用いた分析を実施した[2]。なお，分析を実施するにあたっては，比較の対象となる 2 つのコーパスのどちらかだけに生起する語彙項目が大量に出てくることが予想される。事実，TED Talks においては特定のトークだけに頻出するような語が固有名詞を中心として大量に存在する。こうした事情による影響を最低限にとどめ，より一般的な語彙項目の相対頻度に基づいた分析を実施するため，TED Talks だけに出現する語，および COCA だけに出現する語をすべて除外し，結果として残った 28,180 種のレンマと各コーパス内での頻度を入力値とした[3]。その結果を表 2 に示す。なお, 数値計算と後に示すグラフの出力には R（Version 3.4.3）を使用した[4]。

表 2　TED コーパスと COCA［使用域別］の比較

コーパス	順位相関係数（Kendall's *tau*）	*p* 値
TED ⇔ COCA [Popular Magazine]	0.61	< 0.001
TED ⇔ COCA [Spoken]	0.59	< 0.001
TED ⇔ COCA [Academic Journal]	0.57	< 0.001
TED ⇔ COCA [News]	0.57	< 0.001
TED ⇔ COCA [Fiction]	0.51	< 0.001

最も大きい順位相関係数を示したのは COCA［Spoken］との組み合わせではなく，COCA［Popular Magazine］との組み合わせであった。その後 COCA［Academic Journal］および COCA［News］がほぼ同じ値で続き，最も小さい係数を示したのは COCA［Fiction］であった。TED Talks の言語は基本的に話し言葉ではあるが，いわゆるダイアログではなく，聴衆に向けてなされるプレゼンテーションの言語である。したがって，典型的な話し言葉コーパスには見られない，「書き言葉」的な特徴が生じているとしても不思議ではない。プレゼンテーションでは基本的に，相手の返答によって共有知識の有無や理解度を確認して発話の内容を微調整することは難しい。したがって，情報の自然な流れをあらかじめ考慮した上で，適切な語彙，構文，そしてレトリックを用いた「語

り」を展開することが求められる。これは通常，「書き言葉」に求められるような特徴である。表 2 の結果は，TED コーパスが基本的には話し言葉でありながらも，説得的・説明的なプレゼンテーションの言語であるために，書き言葉の特徴をいくぶん備えていることを示唆している[5]。

上記の観察は TED コーパスの特殊性に着目したものであるが，TED コーパスが英語という言語の基本的（あるいは普遍的）な特性を抽出するのに役立つ資源であることについても触れておきたい。図 8 は TCSE のデータを X 軸，COCA［All］のデータを Y 軸に，レンマごとの頻度をプロットしたものである（Kendall's *tau* = 0.62, $p < 0.001$）[6]。コーパス間で規模が大きく異なるため，プロットの際に対数変換を施しているが，付加された近似曲線の形状からわかるとおり，両コーパスには線形に近い相関がみとめられる。このことは，TED コーパスがプレゼンテーションという固有の使用域に属し，その特徴を多分に備えたデータである一方で，同時に英語という言語の基本的な特徴を保持したものであることを示唆している。

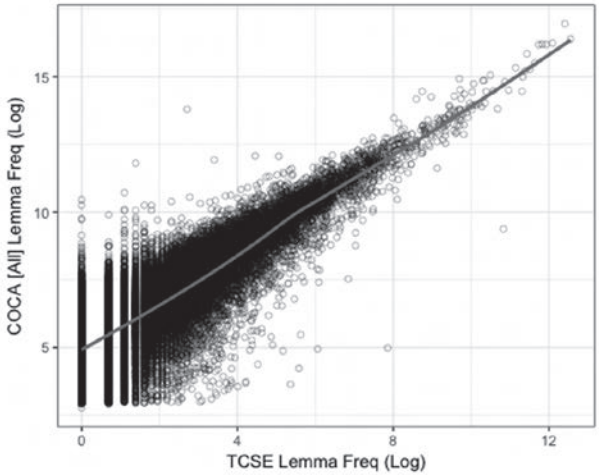


図 8　TCSE と COCA［All］のレンマ相対頻度プロット

以上，本節では TED Talks のコーパスとしての特徴を概観した。次節ではこのような特徴を持つ TCSE が，いくつかの認知言語学上の概念と強く結びついていることを論じる。

4．TCSE の認知言語学的特徴

本節では，TED コーパスを用いた言語研究／言語教育が，認知言語学（cognitive linguistics）の言語観と一致したものとなり得ることについて述べる。具体的には 1）用法基盤モデル，2）注意の枠，3）グラウンディングという 3 つの理論的概念を取り上げ，TCSE との関係について簡潔に論じていく。

4.1 用法基盤モデルの考え方

言語の用法基盤モデル（usage-based model of language）とは，認知言語学が掲げる言語観の基礎をなす考え方であり，言語を現実の使用から切り離された静的な体系として見るのではなく，個々の発話における状況や話者の一人称的体験と結び付いたスキーマ・ネットワークの動的な構造として見なすことを重視した言語観である（cf. Langacker, 2000; Bybee, 2010）。McEnery and Hardie（2011）が強調するように，用法基盤モデルの考え方はコーパス言語学と高い親和性を示す。一般にコーパスは語彙項目や構文の豊富な事例を提供し，言語研究や教育を現実の使用と結びつけると考えられている。しかし，多くのコーパスは豊富な用例・用法を提供するものの，それらが「現実の使用」の側面をどれだけ保持しているかについては考察の余地がある。多くのデータは文脈情報を欠いており，また話者の一人称的視点に基づく発話時の事態認識のあり方を再構成する材料にも乏しい。

一方で TED コーパスの場合はデータのすべてが個々に独立したトークであり，発話者の情報も開示されているため，取り出された表現は，それがいかに断片的であっても，固有の文脈や談話的意図と結びつけることが比較的容易である。もちろん，TED カンファレンスというきわめて限定された発話環境（＝聴衆を前にした 10〜20 分程度の英語によるプレゼンテーション）のみが対象となってしまうが，現実の使用と結びついた言語研究と教育を追求していくための貴重なリファレンスとなり得る。

4.2 注意の枠

注意の枠（windows of attention）という用語は Langacker（2001, 2012）によるもので，Chafe（1994）のイントネーション・ユニット（intonation unit）と共通した概念である。言語学や言語教育において，センテンスという単位は確固たる地位を有しており，多くの研究が，「あらゆる表現はセンテンスごとに

発話・理解されている」という前提のもとになされている。しかし，Langacker によると実際の談話における言語産出や理解は必ずしもその通りではない。言語使用者の注意の枠はより限定されており，多くの場合，言語産出や理解は主要な要素（trajector）とそれに次ぐ重要度を持つ要素（landmark）を中心とした局地的な構図の連続として展開する。このような観点から見ると，センテンスの区切りとは多分に恣意的なものであり，これを過度に重視することは，注意の枠の自然なつながりを適切に捉えることを阻害する。2.3 節で示したように TCSE において基本となっている言語単位はセンテンスではなくセグメントである。セグメントは TED のトランスクリプトが「字幕 1 画面分」を基準に構成されていることを利用した単位であり，形式的なセンテンスの幅にとらわれない。TCSE の提供する言語データ単位が Langacker の論じる注意の枠と完全に一致するわけではないが，話し言葉においてとりわけ重要な「注意の枠に基づいた言語の発話と理解」を考えるにあたり，興味深い材料になると思われる。

4.3 グラウンディング

グラウンディングとは Langacker（2002, 2008）の用語であり，談話の参与者間で共有された知識基盤のもとで事物を認識様態的（epistemically）に同定する認知的プロセスを指す。英語ではあらゆる名詞句（nominal）と定形節（finite clause）はグラウンディングのプロセスを経てはじめて成立する。グラウンディングは明示的なマーカーを伴わずに実現することもあるが，英語の場合，グラウンディング叙述詞（grounding predicate）と呼ばれる特定の語／形態素がこれを起動する役割を担う。名詞に付加する冠詞や，節を構成する際の法助動詞や時制接辞は，グラウンディング叙述詞の典型である。4.1 で論じた用例基盤の考え方とも重なるが，グラウンディングは言葉によって表される内容を現実世界に位置付ける役割を果たす。こうした考え方に基づく言語研究／言語教育を実施するにあたっては，コーパスの豊富な事例が有効であるが，従来的なコーパスでは，必要な文脈的・談話的情報のすべてを再現することが難しい。その点，トークごとに完結しており，かつ発話の状況や文脈情報が完備されている TED コーパスであれば，様々な関連要素を考慮に入れた言語研究／言語教育が可能になる。

5．まとめ

本稿では TED Corpus Search Engine（TCSE）の概要と可能性について論じた。2 節では TCSE の主な機能を概観した。3 節では TCSE に含まれるデータの様々

な統計値と共に，TED Talks の言語コーパスとしての特徴を示した。また，TED Talks の言語が単なる「話し言葉」ではなく，ある種の「書き言葉」的性質を有していること，しかしその一方で，大規模な均衡コーパスが示すような普遍的特徴をある程度備えていることについて述べた。そして 4 節では，1）用例基盤モデル，2）注意の枠，3）グラウンディングという 3 つの理論的概念を取り上げ，TCSE が認知言語学の考え方と一致する特徴を持つことについて論じた。

謝　辞

本稿の内容は英語コーパス学会第 43 回大会（関西学院大学）シンポジウム「話し言葉コーパスの構築と利用」における発表に基づいている。当日の発表とそれに先立つワークショップでは数々の有益なコメントを頂いた。また，『英語コーパス研究』の査読者からは詳細かつ的確な指摘と示唆を頂いた。ここに記して感謝したい。本研究の一部は科学研究費（若手研究 B：25870898）の補助を受けて行われた。

注

1. TED トランスクリプトの翻訳については公式サイトで詳しい手順が公開されている。https://www.ted.com/participate/translate/get-started
2. 順位相関分析に広く用いられている Spearman's *rho* に比べ，Kendall's *tau* はとりわけ同順位の要素が多く存在するデータにおいて高い信頼性を示すと考えられている（cf. Howell, 1997）。
3. 今回の調査では語のレンマと頻度を入力値とし，品詞の区別を行っていない。したがって，品詞が異なっていてもレンマが同形の場合には同じ語として扱われる。これは，TCSE と COCA で用いている品詞解析システムが異なり，正確な比較ができないためである。TCSE で用いている Enju（http://www.nactem.ac.uk/enju）はテキストに Penn Treebank 形式の品詞タグを付与するが，COCA で用いている CLAWS（http://ucrel.lancs.ac.uk/claws）は CLAWS タグセットに基づいた出力を行う。品詞の違いを考慮したより精緻な調査は今後の課題である。
4. 本調査では 2014 年 3 月に取得した COCA の上位 6 万語の頻度データを使用した。COCA の総語数は 2018 年 2 月現在で約 5 億 7 千万語であるが，2014 年当時の総語数は約 4 億 4 千万語であり，当該のデータセットに含まれる上位 6 万語の総頻度は 3 億 7 千万語である。
5. TCSE と COCA［Popular Magazine］が最も高い相関係数を示す理由の 1 つとして、一般雑誌のテキストにはインタビューなど話し言葉の引用が多く含まれることから，話し言葉的特徴が備わっているということも考えられる（匿名の査読者から

の指摘に感謝する）。

6. TCSE と COCA の頻度データをプロットするにあたっては，1 万語や 10 万語などで正規化した調整頻度ではなく，粗頻度を対数変換した値を用いた。これは，各データセットにで極端な値を示す一部の語の影響を抑えるためである。なお，Johannessen and Guevara（2011）によるウェブ・コーパスに関する研究でも同様の手法が採用されている。

参考文献

Bybee, J. (2010) *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Carpenter, P. A. and M. A. Just (1977) “Reading Comprehension as Eyes See It.” In Just, M. A. and P. A. Carpenter (eds.), *Cognitive Processes in Comprehension*. New York: Psychology Press, pp. 109–140.

Chafe, W. L. (1994) *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: University of Chicago Press.

Davies, M. (2010) “The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English.” *Literary and Linguistic Computing 25*, 4: 447–464.

Gray, C. J. (2012) “Readability: A Factor in Student Research?” *The Reference Librarian 53*, 2: 194–205.

Hasebe, Y. (2015) “Design and Implementation of an Online Corpus of Presentation Transcripts of TED Talks.” *Procedia: Social and Behavioral Sciences 198*, 24: 74–182.

Howell, D. C. (1997) Statistical Methods for Psychology. Belmont: Cengage Wadsworth.

Johannessen, J. B. and E. R. Guevara (2011) “What Kind of Corpus is a Web Corpus?” In B. S. Pedersen, G. Nešpore and I. Skadiņa (eds.), *NODALIDA 2011 Conference Proceedings*, pp. 122–129.

Langacker, R. W. (2000) “A Dynamic Usage-based Model.” *Cognitive Linguistics 12*, 2: 143–188.

Langacker, R. W. (2001) “Discourse in Cognitive Grammar.” In Barlow, M. and S. Kemmer (eds.), *Usage-based Models of Language*. Stanford: CSLI, 1–63.

Langacker, R. W. (2002) “Deixis and Subjectivity.” In Brisard, F. (ed.), *Grounding: The Epistemic Footing of Deixis and Reference*. Berlin: Mouton de Gruyter, pp. 1–28.

Langacker, R. W. (2008) *Cognitive Grammar: A Basic Introduction*. *Oxford*: Oxford University Press.［山梨正明（監訳）『認知文法序説』研究社］

Langacker, R. W. (2012) “Elliptic Coordination.” *Cognitive Linguistics 23*, 3: 555–599.

McEnery, T. and A. Hardie (2011) *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.［石川慎一郎（訳）『コーパス言語学：手法・理論・実践』ひつじ書房］

（同志社グローバル・コミュニケーション学部　Email: yhasebe@mail.doshisha.ac.jp）

英語コーパス学会　第43回大会

■第１日目

ワークショップ1【TCSE を用いた TED Talks の全文検索と英語教育への応用】
会　場：関西学院大学（西宮上ケ原キャンパス）第５別館３階　第307教室
日　時：９月30日(土) 10:00–11:30（第５別館１階正面にて9:30受付開始）
講　師：長谷部　陽一郎（同志社大学）
参加費：会員無料。非会員2,000円（当日会員としての大会参加費，２日間共通）。

日　　時　2017年９月30日（土）
受付開始　11:30（第５別館１階正面）
開 会 式　12:30（第５別館１階第２教室）
司　　会　石井　康毅（成城大学）
１．会長挨拶　投野由紀夫（東京外国語大学）
２．開催校挨拶　小菅　正伸（関西学院大学副学長）
３．総会
４．学会賞審査報告　西村　秀夫（三重大学）
５．事務局からの連絡

〈研究発表第1セッション（場所：第５別館１階　第１教室)〉
司　会　佐竹　由帆（駿河台大学）

研究発表１　13:30–14:00
Charles Dickens の The Mystery of Edwin Drood と Thomas Power James によるその続編の文体類似性評価　後藤　克己（中部大学大学院生）

研究発表２　14:05–14:35
TF-IDF を用いた Alice Bradley Sheldon の計量文体分析
木村　美紀（明治大学大学院生）

研究発表３　14:40–15:10
機械学習アプローチによる小説テクストの計量的分析：
― アーサー・コナン・ドイルの作品から ―　黒田　絢香（大阪大学大学院生）

研究発表４　15:15–15:45
Agatha Christie 作品の修辞的特徴に関する分析　土村　成美（大阪大学大学院生）

〈研究発表第２セッション（場所：第５別館１階　第２教室）〉

司　会　　藤原　康弘（名城大学）

研究発表１　　13:30–14:00

Investigating the Impact of Extensive Reading with Data-Driven Learning

Gregory HADLEY（新潟大学）
ハドリー浩美（新潟大学）

研究発表２　　14:05–14:35

日本人英語学習者の関係代名詞の回避
　― CEFR レベルを用いた検証 ―　　高橋　有加（東京外国語大学大学院生）

研究発表３　　14:40–15:10

英語教材としてのアニメ分析　　寺島　芙由（東京外国語大学大学院生）

研究発表４　　15:15–15:45

小中連携に向けた英語授業コーパスデータ構築と
　インタラクション分析の試み　　片桐　徳昭（北海道教育大学）
大橋由紀子（ヤマザキ学園大学）

〈研究発表第３セッション（場所：第５別館３階　第307教室）〉

司　会　　後藤　一章（摂南大学）

研究発表１　　13:30–14:00

複数コーパスの統一的処理を可能にした
　高速コーパスデータベースシステム MyCo の開発　　西村　祐一（元名古屋大学大学院生）

研究発表２　　14:05–14:35

構文構造を活用した学術論文における頻出コリゲーションの抽出

田中　省作（立命館大学）
徳見　道夫（九州大学名誉教授）
宮崎　佳典（静岡大学）
金丸　敏幸（京都大学）
田地野　彰（京都大学）

研究発表３　14:40–15:10
Word2Vec による文学作品の時代比較
―コーパスを軸とした異分野融合研究の試み―　内田　諭（九州大学）
下條　恵子（九州大学）
渡邉　智明（九州大学）
斎藤　新悟（九州大学）
谷口　説男（九州大学）

研究発表4　15:15–15:45
構文情報などを表す木構造の配列による情報処理　田中　省作（立命館大学）
宮崎　佳典（静岡大学）
田辺　利文（福岡大学）
田村　昌彦（立命館大学）

〈休　憩　15:45–16:05〉

〈研究発表第４セッション（場所：第５別館１階　第１教室）〉
司　会　石川　有香（名古屋工業大学）
研究発表１　16:05–16:35
強調語の調査による Popular Music の歌詞の文体研究
渡部　文乃（京都大学大学院生）

研究発表２　16:40–17:10
ホテルのオフィシャルウェブサイトにおける概説文のストラテジー
―Move の構築と分析を中心に―　近藤　雪絵（立命館大学）

研究発表３　17:15–17:45
一般教書演説から見る米国大統領の関心事の変遷
―トピックモデルと時代背景―　木山　直毅（北九州市立大学）

〈研究発表第５セッション（場所：第５別館１階　第２教室）〉
司　会　大谷　直輝（東京外国語大学）
研究発表１　16:05–16:35
【賛助会員発表】コーパスの示す科学的データと学習性・商品性との両立
―『ウィズダム英和辞典』の編集にあたって―　井上　永幸（広島大学）
西垣　浩二（㈱三省堂辞書出版部）

研究発表２　16:40–17:10
英語辞書レーベルとコーパス　田畑　圭介（神戸親和女子大学）

研究発表３　17:15–17:45
怒りを表す類義語と概念メタファー　南澤　佑樹（大阪大学大学院生）

〈**研究発表第６セッション**（場所：第５別館３階　第307教室）〉
司　会　森下　裕三（環太平洋大学）

研究発表１　16:05–16:35
日英対訳コーパス中の「～ことになる」構文とその英訳文間の構造的不一致
大矢　政徳（目白大学）

研究発表２　16:40–17:10
医学研究論文ジャンルにおけるコーパス作成ツール AntCorGen を活用した教育の可能性
— Construction of Corpora for Discipline-Specific Learning
in Medical Research Article Genres　浅野　元子（大阪大学大学院生）

研究発表３　17:15–17:45
Applying Topic Models to Describe a Corpus’s Compositionality:
How can the external criteria be associated with meaningful sets of internal evidence?
Tomoji Tabata (University of Osaka)

《懇親会　時間：18:15–20:30（会場：関学会館　会費：5,000円）》

■第2日目

日　　時　　2017年10月１日（日）
受付開始　　9:30（第５別館１階正面）

ワークショップ2【機械学習を用いたコーパス分析入門】
会　場：関西学院大学(西宮上ケ原キャンパス)第５別館３階　第307教室
日　時：10月１日（日）10:00-11:30
講　師：小林　雄一郎（日本大学）
参加費：会員無料。非会員2,000円（当日会員としての大会参加費，２日間共通)。

〈休　　憩　11:30-12:30〉

講演　　12:30-13:30（第５別館１階　第２教室）
《A Frontier in Learner Corpus Studies: For Better Understanding of L2 Learners》
司　会　　投野由紀夫（東京外国語大学）
講　師　　Shin'ichiro Ishikawa (Kobe University)

〈休　　憩　13:30-13:50〉

シンポジウム　　13:50-15:20（第５別館１階　第２教室）
話し言葉コーパスの構築と利用
司　会　　野口ジュディー(神戸学院大学名誉教授)
The ICNALE：中間言語対照分析の精緻化とアジアにおける学習者コーパス研究の発展を目指して
講　師　　石川慎一郎（神戸大学）
International corpus of Japanese as a second language:
日本語学習者の言語研究と指導のために
講　師　　迫田久美子（広島大学・国立国語研究所）
JECPRESE：JSL と EFL ユーザーのために
講　師　　野口ジュディー(神戸学院大学名誉教授)
TED Corpus Search Engine:
TED Talks を研究と教育に活用するためのプラットフォーム
講　師　　長谷部陽一郎(同志社大学)

閉 会 式　　15:30（第５別館１階　第２教室）
閉会の辞　　井上　永幸（広島大学）

■９月30日（土）
【ワークショップ１】

TCSE を用いた TED Talks の全文検索と英語教育への応用

長谷部陽一郎（同志社大学）

TED Corpus Search Engine（TCSE）は TED が公開している約2,400件の英語プレゼンテーションのトランスクリプトを解析してデータベースに格納し，英語テキストと翻訳テキストの全文検索を可能にした Web システムである。TCSE には英語教育や言語学研究のために TED Talks を役立てるための各種機能が実装されており，本ワークショップでは特に英語教育への応用を念頭においた解説を行う。予定している内容は下記のとおりである。

＜基本編＞
（1） TED Talks と TCSE の概要
（2） 英語と日本語を用いた基本的な事例検索
（3） 教育素材となる Talk を探すために役立つ機能
＜応用編＞
（4） 高度な検索式を使った事例検索と結果の保存
（5）「Pause and Check」機能を活用したリスニング／スピーキング学習
（6） 実践例：TCSE を用いた英語談話標識の指導
＜発展編＞
（7） TCSE の内部構造と理論的背景
（8） TCSE の実験的機能

まず「基本編」では TCSE で何ができるかについて大まかな理解を得ることを目指す。次に「応用編」では主に２つのことを行う。１つは品詞や基本形などの語彙情報を用いてイディオムや構文の事例を効果的に検索する方法を知ること，もう１つは TCSE の「Pause and Check」機能をリスニングやスピーキングの学習や指導に利用する方法を知ることである。最後に「発展編」では TCSE の構造や背景について簡単に解説するとともに，いくつかの実験的な機能について言及する。

本ワークショップは基本的に講義形式で進めていくが，インターネット接続された機器を持参して実際に試していただくのも良いだろう。TCSE は PC（Windows, MacOS）の Web ブラウザ上での使用を基本としているが, 多くの機能はスマートフォンやタブレットでも利用可能である。

■９月30日（土）
【研究発表第１セッション】
【研究発表１】

Charles Dickens の *The Mystery of Edwin Drood* と Thomas Power James による その続編の文体類似性評価

後藤　克己（中部大学大学院生）

米国人 Thomas Power James（以下，T. P. James）は，Charles Dickens の遺作となった *The Mystery of Edwin Drood*（以下，原典）に続編を加えて完全版とし1873年に発表した。この続編は T. P. James がこの続編を “By the Spirit Pen of Charles Dickens, through a Medium.” とアピールしたこともあって大きな論議を呼び，物語性，人物造型の一貫性，言語的側面等の観点から多くの批評がなされた。当時 W. H .B，George F. Gadd，John Cuming Walters 等がそれらの観点から否定的に批評したが，言語的側面への言及は，抽象的・感覚的また部分的なものにとどまっている。そこで原典と続編に，発表時期が1864-5年と原典（1870年）に近い *Our Mutual Friend*（以下，*OMF*）を加えたコーパスを用いて，語彙使用の観点から文体類似性の数量的評価を試みた。

まず，原典，続編および *OMF* に生起する語彙頻度に着目した。作中の発話部は登場人物によって大きく異なるため除外し，作者の本来の文体を最もよく反映していると考えられる地の文のみを用いた。各作品を章単位でサブコーパス化した原典20[1]，続編23および *OMF*67のサブコーパスについて，AntConc を用いて使用語彙のレマでの生起頻度を抽出／構成して語彙頻度表を作成し対応分析を行った。（結果：図１）

つぎに，語彙クラスター (n-gram) には作者の好みが反映されると考えられることから，Mahlberg (2013) で Dickens 作品に特徴的とされている５語クラスター，ならびに類義語 as if / as though を選び，原典，続編および *OMF* での生起頻度を比較した。なお, ここではいずれの作品とも発話部を含むフルテキストを使用した。（結果:図２）

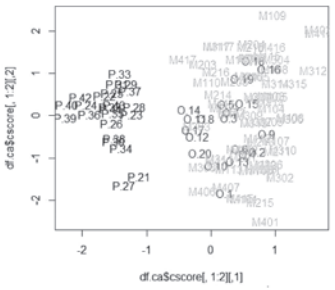


図１　対応分析による1―2軸の列スコア[2]

５語クラスターグループ	頻度計（10万語換算）		
	原典	続編	*OMF*
Body Part	6.3	0.8	8.9
As If	8.5	0.8	6.4
計	14.8	1.6	15.3

２語クラスター	頻度計（10万語換算）		
	原典	続編	*OMF*
as if	112	15	120
as though	29	92	11
計	141	107	131

図２　クラスターの頻度比較

[1] Dickens の原典は23章で構成されているが, T. P. James による完全版の原典部分は, 一部の章が合体して20章構成となっている。
[2] いずれかの作品に70%以上偏在する語，人称代名詞および人名を除き，計 50Tokens 以上の上位547語による。

その結果，原典（O.xx）と *OMF*（Mxxx）のサブコーパスは，まとまった一つのクラスターを形成するものの，続編（P.xx）のそれは明らかに別クラスターを形成しており，また Body Part と As If グループの５語クラスターの続編での生起頻度は原典／*OMF* より圧倒的に少なく，また，as if / as though の生起傾向は，原典／*OMF* と続編とで逆転していることが判明した。これらの結果は語彙使用面で，続編と Dickens 作品との文体類似性が低いことを示しており，続編を “By the Spirit Pen of Charles Dickens, through a Medium.” とする T. P. James のアピールへの懐疑的見方を支持するものでもある。

【研究発表２】

TF-IDF を用いた Alice Bradley Sheldon の計量文体分析

木村　美紀（明治大学大学院生）

本研究では，コーパス言語学の一分野である計量文体論の手法を用いて Alice Bradley Sheldon 作品群に関する文体調査を行う。Alice Bradley Sheldon（1915-1987: 米国）はデビューから約10年間正体不明･性別不明の作家 James Tiptree, Jr. として著作活動を行ってきた作家である。この作家の文体は文芸批評において，Ernest Hemingway と比較されることが多い。特に Silverberg（1975）では “Hemingway was a deeper and trickier writer than he pretended to be; so too with Tiptree, who conceals behind an aw-shucks artlessness an astonishing skill for shaping scenes and misdirecting readers into unexpected abysses of experience. And there is, too, that prevailing masculinity about both of them.” というようにいささか主観的・印象論的に Alice Sheldon の男性名義である James Tiptree, Jr. 作品群と Ernest Hemingway 作品の文体比較が行われている。

計量文体論の手法を用いた Alice Bradley Sheldon 作品群の文体比較では，Alice Bradley Sheldon 全72作品を収録したコーパス（延べ865,802語）と Ernest Hemingway 69作品を収録したコーパス（延べ271,475語）を用いて，Burrows（1987）に基づき高頻度語彙上位50語を指標としてランダムフォレストを行った。その結果，分類正確率は92.20% であり基準値51.06% を有意に上回っているということが判明した。ここから，文芸批評で主張されているようなこの２著者における文体の類似性はとらえることができなかったといえる。

しかしながら，この文体比較は，サンプルサイズが小さいという問題点や，ジャンルの相違などの様々な交絡因子が存在するという問題点がある。このため，発表ではこれらの問題点を克服するために Alice Bradley Sheldon と同時代・同ジャンルで活躍した作家のコーパスを構築して文体比較を進める。具体的には，Alice Sheldon と親交のあった Ursula K. Le Guin の45テキストを含むコーパス（延べ589,481語），Octavia E. Butler の93テキストを含むコーパス（延べ867,396語），Arthur C. Clarke の

104テキストを含むコーパス（延べ467,983語），Theodore Sturgeon の222テキストを含むコーパス（延べ1,777,561語）を構築した。サンプルサイズの違いの影響を減じるため，Theodore Sturgeon 全222作品からランダムに70作品抽出したコーパス（延べ475,704語）を分析に使用した。

分析には，Baayen（2008）で分類感度の高さが指摘されているサポートベクターマシン（SVM）と，金・村上（2007）において分類感度の高さが指摘されているランダムフォレストという統計手法を採用した。指標としては，定義や実行の容易さとその実績から単語頻度，単語の bigram，trigram，小町（2016）や小林（2016）などの情報処理の分野で使用されている TF-IDF の unigram，bigram，trigram という６種類を用いて分類正確率の変化を検証する。６種類のコーパスでのランダムフォレストを用いたテキストの分類正確率はそれぞれ，86.31%，77.70%，66.45%，85.43%，77.26%，62.03% だった。また，SVM を用いたテキストの分類正確率はそれぞれ，99.56%，97.35%，90.29%，99.56%，97.57%，81.23% だった。本発表では，TF-IDF という情報処理の分野で広く使用されている重みづけの概念を言語研究の場に導入し，これらの指標を用いた分類正確率の比較提示をおこなう。また，分類正確率の提示だけではなく，作品の類似度の視覚化を行い，コーパスを用いた計量的研究が文芸批評で行われている研究へどのように貢献できるのかを提示する。

【研究発表３】

機械学習アプローチによる小説テクストの計量的分析
― アーサー・コナン・ドイルの作品から ―

黒田　絢香（大阪大学大学院生）

Arthur Conan Doyle はシャーロック・ホームズシリーズの著者として広く知られる作家であるが，自身の本分と考え注力していた歴史小説はこれまであまり批評や研究の対象となっていない。また，既存の研究はいわゆる ‘close reading’ のアプローチが主体で，対立概念である ‘distant reading’（Moretti, 2000），つまり客観的な量的データに基づいて分析するアプローチは少ない。そこで本研究は，Doyle の推理小説と歴史小説を対象として，語彙頻度や生起パターンなど量的な観点から考察を行うことで，文学研究に新たな視点を提案することを目的とする。リサーチクエスチョンは以下の二つである。

1）Doyle の推理小説群と歴史小説群は統計的手法に基づき機械的に分類することは可能か，またどれほどの精度で行うことができるか。
2）それぞれのジャンルを特徴づける語や表現は何か。

分析対象は，Doyle による推理小説７作品と歴史小説９作品の計16作品である。分類には，アンサンブル学習のアルゴリズムである Random Forests（Breiman, 2001）を

用いた。分類ののち，Tabata（2015）で提案されている Random Forests による分類に寄与した語に注目する特徴語抽出法を参考に，各ジャンルの特徴語をリスト化した。また，単語単体だけではなく，関連性のある単語をグループ化した『トピック』にもジャンル間の差異が現れるのではないかと考え，潜在的ディリクレ配分法（LDA）（Blei et al., 2003）に基づくトピックモデリングを行った。MALLET という java ベースのツールキットを用いて，トピックごとの単語の生起確率，テクストごとのトピックの生起確率を算出し，得られた結果をネットワークグラフに可視化した。

Random Forests による分類の精度は96.58% で，十分に高い値で分類することが可能であった。また，推理小説側の中で部分的に歴史小説と誤分類されたファイルは，推理や謎解きではなく過去の出来事を描写する箇所で，この 'retrospective narrative' が歴史小説と似た語彙パターンを持っているのではないかと考察した。分類に寄与した特徴語には，歴史小説側は 'cried' や 'fight'，また 'head', 'faces', 'arms', 'eyes' のような体の部位を表す語，一方で推理小説側は 'case' や 'found' など捜査に関する語や，'house', 'chair', 'room' など家や家具に関する語が抽出された。

トピックモデリングの結果も Random Forests の結果とある程度一致し，推理小説側に頻出する『犯罪捜査トピック』や『家・家具トピック』などを発見した。設定するトピック数を変化させて実験を行い，最も効果的にジャンル間の差異を発見できるトピック数設定について考察した。

二つの機械学習に基づく手法から得られた結果を総合し，量的な観点からそれぞれのジャンルを特徴づける語やトピックを発見した。

【研究発表４】

Agatha Christie 作品の修辞的特徴に関する分析

土村　成美（大阪大学大学院生）

本研究では，Agatha Christie 作品の文体的特徴について，彼女と同時代に活躍したミステリー作家 Dorothy Sayers との比較を通して分析を行う。文学作品の文体に関する計量的な研究では主に語の出現頻度が分析の指標として用いられており，Christie 作品に関する計量的な分析も語彙を指標とした分析が多い（Lancashire & Hirst, 2009; 稲木，2009；稲木，2013）。テキストに修辞的アノテーションを施すことは処理が複雑になるため，修辞的項目を指標とした分析は語彙や POS タグ・統語構造を指標とした分析ほどには多く行われていない。そのため，本研究ではテキストに修辞的アノテーションを行い，修辞的項目を指標として計量的分析を行う。修辞的項目を変数とした２作家の作品の分類実験を行い，Christie に特徴的な修辞的表現の抽出・検討を行う。

使用データは，Christie の221作品（5,071,282語），Sayers の55作品（1,375,645語）を用いる。両作家で作品数・総語数共に大きく異なるため，それぞれの作家から50作品を無作為抽出し，分析に用いた。

修辞的アノテーションには DocuScope を使用した。DocuScope は Kaufer & Butler（1996）におけるレトリック理論と，Kaufer & Butler（2000）における言語表象理論を基礎として構築されたテキスト分析・視覚化のためのツールであり，言語表現を101の Language Action Types（LATs）に分類し，タグ付けを行う。

DocuScope を用いて得られた作品ごとの LATs の頻度情報を指標として，機械学習の一種 Random Forests（Breiman, 2001）を使用した分類を行った。分類の正解率は92%～94% であった。Random Forests の分類において寄与率の高い項目について検討を行う。

寄与率の高い項目の中で，Sayers 作品と比較して Christie 作品において最も過剰使用されているのは，一人称代名詞と動詞や前置詞の組み合わせ（*I think, I feel, for me* など）で一人称としての意識を表す Self-Disclosure であった。推理小説故に会話文でこの LAT が多く出現することは明らかであるが，Sayers と比較すると Christie 作品では地の文でも一定数使用されていることが確認された。Christie 作品は作中の登場人物が語り手を務める作品も多く，両作家の地の文での語り手の違いが反映されている可能性があると考えられる。また Christie は熱中や傾倒を表す Intensity（*very, indeed, I do, great* など）の使用も多いことが明らかになった。特に Intensity に関しては同一作品における同じ語の繰り返しが顕著に見られ，Christie の語彙多様性が晩年になるに従い低下したこと（Lancashire & Hirst, 2009; Le et al., 2011）を反映していると考えられる。以上のような修辞的特徴から，Christie 作品の文体的特徴の分析を試みる。

■９月30日（土）

【研究発表第２セッション】

【研究発表１】

Investigating the Impact of Extensive Reading with Data-Driven Learning

Gregory HADLEY（新潟大学）・ハドリー浩美（新潟大学）

This presentation discusses an ongoing research projected investigating the use of data-driven learning (DDL) as a means of stimulating greater lexicogrammatical knowledge and reading speed among lower proficiency learners in an extensive reading program. From April 2015 to July 2017, students from six extensive reading classes were chosen for this study. For 16 weekly 90-minute sessions, an experimental group (21 students) used DDL materials created from a corpus developed from the Oxford Bookworms Graded Readers, which contained 186 books from all seven levels with a total of 1,715,160 tokens (17,670 word

types). The control group (28 students) had no DDL input. All students were required to read a minimum of 200,000 words during the course. Students not reaching the 200,000 word threshold were removed from this study. Quantitative data from a C-test (Klein-Braley & Raatz, 1984) constructed from an upper-level Bookworms reader. A speed reading test by Quinn, Nation, & Millett (2007) was also selected. A pre-test post-test design was used, and dependent as well as independent samples *t* tests were used. Pre-test analysis found that the experimental group was in statistically distinct from the control group in terms of having lower levels of second language proficiency. Post-test scores found that, within both groups, the learners improved significantly, with high impact factors. However, the experimental group improved more to the point that the entered the same statistical bands as the control group. Post-test findings also indicate that students using the DDL materials were reading more books and reading faster than the control group. The study concludes that an informed use of DDL can work with lower proficiency learners, and that the methodology can be used to improve receptive learning and lexicogrammatical proficiency better than extensive reading alone.

【研究発表２】

日本人英語学習者の関係代名詞の回避 ― CEFR レベルを用いた検証 ―

高橋　有加（東京外国語大学大学院生）

本研究では，同一実験参加者による①通常の英作文と，②関係詞を使うように指示のある英作文の，２つのタスク内での関係詞の使用頻度とエラー率を比較することにより，知識はあっても使用が回避される傾向が CEFR レベル別にどの程度あるのかを明らかにすることを目的とする。

関係詞は日本人英語学習者にとって難しいとされるものの１つであり，使用が義務的でない文法項目である。Shachter（1974）は，学習者にとって難しいものは使用が回避される傾向があるとし，その例として関係詞を取り上げている。また，近年日本の英語教育にも大きな影響を及ぼしているヨーロッパ言語共通参照枠（Common European Framework of Reference: CEFR）の６つのレベルを同定する言語特性である基準特性（criterial features: Hawkins & Filipović, 2012）を学習者コーパスから抽出する研究が行われており，Hawkins（2009）は基準特性として有効な文法項目の１つとして関係詞の使用を挙げている。しかし，自発的な産出が求められるパフォーマンステストなどでは，知識があっても実際に産出されないと評価することができないため，自然に産出される関係詞と，意識的な関係詞の使用を求められるタスク内における頻度及びエラーについて比較した。研究設問は以下のように設定した。

RQ1. 関係詞の回避または不使用の現象が実際にどのくらいあるのか？
RQ2. 意図的に関係詞を使用するようにした場合，エラー率は増加するのか？

対象とする関係詞は，that，which，who，whose，whom の５つで，省略形は含まない。また，SLA 理論の習得難易度階層を示す Noun Phrase Accessibility Hierarchy（NPAH: Keenan & Comrie, 1977），SO Hierarchy Hypothesis（SOHH: Hamilton, 1994）の分類ごとにも集計した。実験参加者は，英検級を保持する日本人英語学習者93名である。文部科学省（2015）の対照表によると，５−３級：A1，準２級：A2，２級：B1，準１級：B2，１級：C1 であるため，それぞれのレベルから20名程度の実験参加者を集めた。

データ収集方法として，全ての実験参加者に２種類の英作文タスク（20分間辞書なし）を異なる日時に受験してもらった。１つ目は通常の自由英作文（描写タスク）で，２つ目は同一のタスクに「できるだけ関係詞を使うように」という指示を加えたタスクである。データ処理として，全ての英作文を書き起こし，(a) 関係詞の表層形，(b) NPAH，SOHH のタイプ，(c) エラー情報について，１文ごとに人手でアノテーションを施した。

調査の結果，関係詞を使うように指示のあるタスクでは，全レベルで関係詞の頻度が大幅に増加した。このことから，関係詞の知識があっても間違いを恐れたり，関係詞の使用が不必要であると判断した場合には使用されない関係詞が多くあることが分かった。エラー率に関しては，A2 レベルにおいて１回目より２回目の英作文内に特にエラーが多く見られたことから，間違いを恐れるために関係詞の使用を避ける傾向がある可能性が示唆された。

【研究発表３】

英語教材としてのアニメ分析

寺島　芙由（東京外国語大学大学院生）

アニメ人気の高まりを受け，アニメを英語教材として提案する研究（佐々木，2005）や，英語版アニメを活用した授業（吉田，2012;佐藤，2014）がみられる一方，実際にアニメが効果的な学習方法かどうか検討を行っている研究は稀である。そこで本研究では，アニメの英語学習への活用の可能性を，コーパスを使って語彙の観点から調査した。

対象は英語に吹き替えられた日本のアニメ４作品で，比較対象としてアメリカで制作された映画１本も分析した。以下，３つの研究設問に沿って，調査の手法と結果を示す。

1．アニメの視聴は学習に適しているか。

アニメ全てのテキストと，BNC spoken corpus の上位100語および200語，バイグラムおよびトライグラムの上位100件を比較したところ，その多くがアニメに高頻度で現れたことから，アニメは効果的なインプットである可能性が示された。ただし，AntWordProfiler を用いて分析を行ったところ，General Service List（GSL）と Academic Word List（AWL）に含まれる語彙がアニメのテキストの約90% を占めていたため，最低でもこれらの語彙知識がない初級・中級学習者にはアニメによる英語学習は適切ではない可能性が示された。

2．アニメごとにどのような特徴があるか。

分析結果を以下の表に示す（括弧内は割合（%））。

	時間	延べ語数	TTR	GSL 1000, 2000	AWL	その他
アニメ1	110分	14291	0.163	12478（87.31）	294（2.06）	1519（10.63）
アニメ2	110分	11459	0.127	10390（90.67）	107（0.93）	962（8.4）
アニメ3	110分	13062	0.145	11642（89.13）	207（1.58）	1213（9.29）
アニメ4	110分	11419	0.131	10268（89.92）	126（1.1）	1025（8.98）
映画	124分	7748	0.197	6747（87.08）	149（1.92）	852（11）

簡単な語や同じ語の繰り返しが多いアニメ２は，難解な語を多く使い，延べ語数や同じ語の繰り返しが少ない映画と比べ，学習者にとってより効果的である可能性などが示された。

3．アニメによる自主的な学習のためにどのような援助を行うのが効果的か。

AntConc の keyword list を使ってアニメ２を分析すると，アニメに特有な単語として，固有名詞やアニメにおける用語などが発見された。加えて，アニメの内容に基づいてとりわけ多くみられるコロケーション（例：wish for）が存在することがわかった。これらの単語や表現を事前に提示することで，学習者の負担を減らし，学習を促すような援助方法が考えられる。

発表では，本研究の課題や，今後必要とされている語彙以外の面での研究も考察する。

【研究発表４】

小中連携に向けた英語授業コーパスデータ構築とインタラクション分析の試み

片桐　徳昭（北海道教育大学）・大橋由紀子（ヤマザキ学園大学）

1．背景と目的

平成32年（2020年）度から新学習指導要領の完全実施となり，小学校で英語が教科化される。文科省が謳う小中連携の一つに，小中9年間の学びの中の「学習指導の

継続性」という視点がある。そこで本研究では，小中学校での英語授業のコーパスの構築において，インタラクションタグ付与が小中の学習指導の継続性を調べる上で有用なアノテーションとなるかの調査を試みた結果の報告をする。以下の研究課題に基づき，英語の授業内でのインタラクション分析を行い，小中の接続点である，小学校６年生の終了時と中学１年生の開始時の英語授業の「継続性」がどのように観察されるかについて調査した。

研究課題１．インタラクション情報は小中英語授業の「継続性」分析に有用なアノテーションか。

研究課題２．連続する小中英語授業の「継続性」はインタラクションに観察されるのか。

２．データ収集と分析方法

同じ国立大学に附属する小学校６年生の最後の英語の授業４回と連続する年度の中学校１年生の最初授業６回の授業データを収録した。計10回の授業データについて教師・生徒の発話（日英両語）について，話者タグ，言語タグに加えてインタラクションタグを付与した。インタラクションタグとして Walsh（2006）の言う SETT（Self-Evaluation of Teacher Talk）から４つの interaction mode（managerial, materials, classroom context, skills and systems）と Ellis（1984）の述べる social という考えを属性値として組み込んで分類した。インタラクションタグは主に発話ターンごとに付与したが，質的変化があると判断された場合には，同１ターンを細分割してインタラタクションタグを付与した。

３．結果

インタラタクションタグ数は小学校100回（*M*=25.0）・中学校136回（*M*=22.7）となり，小学校・中学校ともに授業体制を構築する managerial mode が最頻出（小学51回［*M*=12.8］，中学74回［*M*=12.3］）となり，教師主導の授業展開の傾向が見られた。また，skills and systems mode による言語材料の習熟活動（小学26回［*M*=6.5］，中学33回［*M*=5.5］），classroom context mode による教師生徒間のインタラクション（小学17回［*M*=4.3］，中学18回［*M*=3.0］）も同等の傾向を示した。しかし，授業目標の核となるコミュニケーション活動を示す materials mode において，小学４回［*M*=1.0］，中学13回［*M*=2.2］となり中学校ではわずかながら発展性が示唆される結果が観察された。

発表当日は，（1）コーパス構築の手順，（2）データ整理の方法，（3）分析結果の詳細ついて述べ，（4）インタラクションタグ付与に関して Walsh（2006, pp. 82-91）が言っている mode switching, mode side sequences, mode divergence といった deviant cases の問題についても触れ，データの利用の拡張性について考察を加える。

■９月30日（土）
【研究発表第３セッション】
【研究発表１】

複数コーパスの統一的処理を可能にした 高速コーパスデータベースシステム MyCo の開発

西村　祐一（元名古屋大学大学院生）

記録方式の異なる様々なコーパスが併存している今日，複数のコーパスを統一化された環境で高速に処理できるデータベースシステムを開発することの意義は大きい。そこで，リレーショナルデータベース管理システム MySQL を利用して，英語コーパスのデータベースシステム（MyCo と呼称）を開発することとした。例えば，BNC と COCA は設計が大きく異なるが，MyCo を用いれば，これらを同一の利用環境で，選択的にあるいは統合して処理することが可能である。随時，様々なコーパスを追加できる設計にしたことによって，入手した複数のコーパスを比較利用できる点も研究上，有益である。また，データ処理にかかる時間を大幅に短縮できたことも，研究遂行上，極めて重要な点である。本発表では，コーパス利用の観点から MyCo の特長を述べる。

１．データベースに記録するデータ

最小単位は WLP（Word, Lemma, POS-tag）である。研究目的に応じて W, L, P を組合せた検索式を指定してデータを抽出する。

２．データ抽出処理

コーパス利用の中心作業は，検索する語（またはレマ）とその周辺に共起する要素の文字列を抽出することである。MyCo では抽出範囲を検索語の前後各10要素に固定し，検索語の WLP をノードとする21個の WLP から成る文字列（21gram）をテキストファイルに出力する。さらに，このテキストファイルを利用者が直接利用することも可能である。

３．資料の加工

MyCo は，抽出した 21gram をもとにピクチャー，kwic リスト，コロケーションリストを作成できる。ピクチャーを例に説明すると，表示範囲，集計対象（Word, Lemma, POS），値（頻度，MI-score，t-score），特定の位置に現れる語またはレマを含むテキストの一覧表示，などである。

４．検索処理時間

BNC によるデータベースを例に，because（頻度100,509），something（頻度50,062），maybe（頻度10,025）について 21gram 抽出時間の実測値を例示すると，それぞれ2.7秒，1.3秒，0.3秒である。１億語規模のコーパスから頻度５万件程度の語の 21gram をほぼ１秒で得ることができ，高速レスポンスを実現している。これは，研究上，全くス

トレスを感じずに済む速度である。また，単一の語またはレマの検索に加えて，イディオム kick the bucket のような複数要素を組み合わせた検索も可能である。

５．実行環境

MyCo を，Linux（CentOS 6.9），Perl v.5.10，MySQL v.5.1 で開発し，利用している。同等環境の PC であれば MyCo を容易に搭載できる。

【研究発表２】

構文構造を活用した学術論文における頻出コリゲーションの抽出

田中　省作（立命館大学）・徳見　道夫（九州大学）・宮崎　佳典（静岡大学）
金丸　敏幸（京都大学）・田地野　彰（京都大学）

現在，分野別の学術論文コーパスをもとにした，分野に依存しない頻出表現の整備を試みている。コーパスからの頻出表現の抽出には，論文コーパス内の各英文で，分野依存性の高い語を適当な痕跡や品詞に置き換え，n-gram で計数すれば良いように考えられる。しかし，n-gram は，“take 〜 into account” のように自由項を挟んだ不連続な頻出表現の捕捉が難しく，また適切な n の設定も問題となる。そこで，本研究では，痕跡に相当する情報を含む英文から，構文構造を考慮した頻出表現の抽出法を提案する。構文構造の考慮と n-gram の重み付けによって，痕跡が木構造における節表示となったコリゲーションの抽出も可能となる。

提案手法は次の通りである。まず，痕跡の取り扱いである。構文構造を参照し，内容語を含まない，痕跡をまとめて導出する最上位の節表示に置換する。たとえば，“The algorithmic procedure takes supremum norm into account” という情報系論文の英文で，“algorithmic”, “supremum”, “norm” を分野依存性の高い語とし，痕跡に相当したとする。SNLPG（n.d.）で与えられる構文構造を参照し，痕跡を節表示に置換すると，“<NP> procedure takes <NP> into account” となる。ただし，<α> は α という節表示，NP は名詞句を表す。コーパス内の英文を全てこのように置き換える。

次に，このように節表示を含んだ英文の n-gram の重み付け計数である。節表示を含む n-gram の重みはそれが元の英文で内含している語の n-gram 数とする。例えば，n=4 のとき “<NP> procedure takes <NP>” の重みは 3，“takes <NP> into account” の重みは 2 と計数される。高次の節表示を多く含む n-gram の方が重く計数される傾向となる。

京大論文コーパスに対して，分野依存性の高い語を痕跡に見立て，上記方法で n を適当に動かし，累積的に計数した。その結果，“between <NP> and <NP>” や “according to <NP>” という具合に節表示を含む，より分かりやすい形で，様々な長さの頻出表現を抽出できることを確認した。

【研究発表３】

Word2Vec による文学作品の時代比較
― コーパスを軸とした異分野融合研究の試み ―

内田　諭（九州大学）・下條　恵子（九州大学）・渡邉　智明（九州大学）
斎藤　新悟（九州大学）・谷口　説男（九州大学）

近年，コンピュータの発達と言語データの蓄積により，自然言語処理の技術は大幅に向上してきている。ニューラルネットワークによる翻訳精度の向上（Wu et al., 2016），単語行列をベクトル化して単語の意味を数値で表す Word Embedding の手法（Levy and Goldberg, 2014）の発展など，その進歩は目覚ましく，大きな注目を集めている。一方でこれらの手法を言語学や文学，文体論などのいわゆる人文社会系の研究への応用については，十分に議論されているとは言えず，未開拓の部分が多く残されている。

本研究の目的は，Word Embedding の代表的な手法の一つである Word2Vec（Mikolov et al., 2013）を用いて単語の用法を独自に作成した文学コーパスで検証し，文学研究への応用の可能性を探ることである。文学作品について計量的あるいは統計的なアプローチを用いる研究は増加傾向にあるが，Word Embedding の手法を用いた研究は限られており，その応用方法については現段階では未知数であるといえる。この手法の最大の特徴は，共起関係に代表される syntagmatic な関係にある単語ではなく，類似のベクトルを持つ単語を探索することで paradigmatic な関係にある単語を検証することができるという点である。特定の単語と類義的に使用されている単語を調査することで，その単語が使用される文脈や含意などを明らかにすることが可能となる。

本研究では1900年代前半の文学作品からなる「前半コーパス」（約180万語）と1900年代後半からなる「後半コーパス」（約171万語）を作成した。対象となる作品については，アメリカ文学作品の中から，金融業界を取り扱った作品及び現実を写実的に描いたとされるリアリズム小説を中心に選定した。20世紀のアメリカは1945年の第二次大戦終戦を機に国際政治の舞台で超大国の地位を獲得しただけでなく，経済面でも規制緩和を繰り返して活性化を図るなど，社会的に大きな変化を経験している。そしてこのような変化は文学作家たちの意識形成に影響を及ぼしており，作品内容だけでなく語彙レベルでその変化が生じていると考えられる。

これらの２つのコーパスを入力としてそれぞれの Word2Vec のモデルを構築した。その後，単語の頻度表から高頻度の名詞を選定してコサイン類似度から同義的に使用されている単語のリストを生成した。その結果，類義語のリストは年代によって興味深い違いを示すことが明らかになった。例えば，money は「前半コーパス」では pay, work, dollars などと類似度が高い。一方，「後半コーパス」では interest, market, bonds などが類似度の高い語としてリストされた。これは「前半コーパス」の時代には労働に対する具体的対価として金銭が語られているのに対し，「後半コーパス」で

は金融商品など利益を生む無形の財やサービスとして語られているということを示唆する。これらの実験結果に対して，本研究では文学・政治学等の観点からの解釈を試みる。言語データの裏側にある事実や社会変化などを多角的に読み解き，コーパスを基軸とした異分野融合の可能性についても議論を行う。

【研究発表４】

構文情報などを表す木構造の配列による情報処理

田中　省作（立命館大学）・宮崎　佳典（静岡大学）
田辺　利文（福岡大学）・田村　昌彦（立命館大学）

近年，実用的な構文解析ツールも増え，構文構造などの語や品詞の連鎖よりも高次の言語情報処理が可能となり，コーパス研究においてもその活用が期待される。本発表は，このような構文情報付き言語データをコーパス研究で活用するための技術的提案である。

コンコーダンサ等の既存分析ツールを超える処理をする際，文字列・語列程度の取り扱いであれば自身でプログラミングするコーパス研究者も少なくない。一方，構文情報は複雑で「木構造」とよばれるデータ構造で表されることが多く，その取り扱いは格段に難しくなる。木構造は「構造体」「ポインタ」「再帰」など，語列処理では無縁のプログラミング構成概念が求められる（情報処理推進機構，2016）。情報系学生でも慣れないうちは混乱することもあり，コーパス研究者にとってはなおさらである。そこで，本発表ではプログラミングの初期に学習し，語列処理等でも使う「配列」，具体的には２つの配列 c, d で木構造を表し，処理する方法を提案する。

配列の一つ c は句表示や語といった節・葉の言語情報，もう一つの配列 d は根（頂点の節）からのパス長を格納する。格納順序は，木構造を根から深さ優先最左探索した順序である。配列では木構造が平坦化してしまったようにみえるかもしれないが，次の基準で配列を先頭から走査し，節 i を解釈すると，木構造が表現されていることがわかる。

１．i=0 ⇒ 根
２．d[i]=d[i-1]+1 ⇒ i は，i-1 の最右子節
３．１，２以外 ⇒ i は，j<i で d[j]=d[i]-1 となる最も大きな j の最右子節

なお，d[i] は，配列 d の i 番目の要素を表し，最右子節は i-1 までに導出される子節のなかで最も右側の子節という意味である。

このように木構造を配列に直すことによって，木構造間の照合を比較的単純な配列間の照合に帰着でき，語列処理とさほど変わらない複雑さで実装できる。本発表では，構文情報が付された英文に対して，構文的な特徴を考慮した，本方式による検索事例も紹介する。

■９月30日（土）
【研究発表第４セッション】
【研究発表１】

強調語の調査による Popular Music の歌詞の文体研究

渡部　文乃（京都大学大学院生）

１．背景・目的

本研究の目的は，Popular Music の歌詞の文体を明らかにすることである。Popular Music とは，不特定多数の聴衆に向けられた，利益を目的とする音楽（cf. Tagg 1982: 41-42）のことで，その起源は19世紀末にも遡る（cf. Frith 1986: 79）ことができる。しかし，Popular Music が本格的に学術分野として扱われ始めたのは最近のことであり（Tagg 1982: 37），歌詞の研究はほとんどない。文体は，社会言語学や歴史言語学などの様々な言語研究の議論において考慮に入れるべき重要な事項であるため，本研究はそのような研究を行う前段階として，Popular Music の歌詞の文体について調査を行う。

２．方法

本研究は，強調語の分布に注目する。強調語を研究対象とした理由は，この項目の調査によって，① 話し手と聞き手との関わり（cf. Hyland 1998），② 文体のフォーマリティー（cf. Yaguchi et al. 2009），③ 主観的性（cf. Hyland 1998）などの文体的特徴が明らかになると期待されるからである。本研究では，申請者が構築した American Popular Music Corpus of English（PMCE-US）という英語歌詞コーパスを使用し，20個のジャンルから構成される均衡コーパス Manually Annotated Sub-Corpus（MASC）と比較して調査するため，Popular Music の歌詞の④ジャンル間における位置づけについても言及する。本研究は形容詞を修飾する副詞の強調語をすべて抽出し，頻度をジャンルごとに比較した。

３．結果・考察

PMCE-US と MASC の比較から明らかになったのは，歌詞が独特の文体をもつということである。強調語全体の頻度に関して，PMCE-US には200個以上の強調語が出現したが，この頻度は MASC の中で最も強調語の出現頻度が高いジャンルである twitter と比べても著しかった。さらに強調語の種類に関しても，幅広いジャンル（e.g. journal, court, e-mail）で見られる very や，口語のジャンル（e.g. face-to-face）や新しいジャンル（e.g. blog）に見られる really はほとんど現れず，出現数の８割を占めていたのは so であった。

このような結果から，歌詞は MASC のどのジャンル以上に，話し手が積極的に聞き手を説得する（＝積極的に聞き手と関わろうとする），主観的でインフォーマルな文体であることが明らかになった。また，強調語 so が歌詞の強調語の全体頻度の８割を占めていたのは，音楽のリズムが短い音節の語を好む傾向があることや，最も収益が見込まれる若者世代の言葉に近づけるための作詞者の意図的な言葉の選択と

考えられ，歌詞はそのように他のジャンルでは見られない特徴をもつジャンルであることが分かった。

【研究発表２】

ホテルのオフィシャルウェブサイトにおける概説文のストラテジー
― Move の構築と分析を中心に ―

近藤　雪絵（立命館大学）

本研究はロンドンのホテルのオフィシャルウェブサイトに掲載された概説文を Swales（1990, 2004）が提唱したジャンル分析の手法を用いて分析し，読み手にアピールするストラテジーを探求することを目的としたものである。ロンドンの３－５つ星のホテルのウェブサイトに掲載された概説文（３つ星ホテル11，４つ星ホテル66，５つ星ホテル47の計124）を集積し，書き手の意図と特徴的な言語表現を元に分類したところ，３つの Move と３つの Step が構築された。Move と Step の概要を表１．に，ホテルのグレード（星）別 Move 採択率を表２．に示した。

表１．Move・Step とその機能

Move		機能
Move 1	Defining self	ホテル自身を定義する
Move 2	Establishing features	ホテルの特徴を確立する
	Step 1: Describing the history/architecture	歴史／建築を述べる
	Step 2: Describing the location	所在地を述べる
	Step 3: Describing the facilities	設備を述べる
Move 3	Establishing connections	ホテルと読み手との関係を築く

表２．ホテルのグレード別 Move 採択率

	3-star	4-star	5-star
Move 1	81.8%	83.3%	83.0%
Move 2	100.0%	92.4%	87.2%
Move 3	81.8%	75.8%	53.2%

Move 1, 2 は採択率が全てのグレードにおいて８割を超えており，自身を定義付けてから特徴を確立することがホテルの概説文の典型パターンであることがわかった。ホテルのグレードが下がると Move 2 の採択率は高まり，3-star では全ての概説文に Move 2 が採択された。一方で，Move 3 は 3-star ホテルでは Move 1 と同じ８割強の採択率であるが，5-star ホテルでは５割強にとどまった。Move 3 では二人称代名詞を用いて読み手に呼びかけたり，予約を促したりする表現が見られ，中グレードホテル

ではこのような表現を使い読み手と関係を築くストラテジーが使われていた。概説文は一見すると自由に創作された文章のように思われるが，Move 分析を行うことで典型的なパターンが存在し，中グレードホテルと高グレードで読み手にアピールするストラテジーが異なることがわかった。

【研究発表３】

一般教書演説から見る米国大統領の関心事の変遷 ― トピックモデルと時代背景 ―

木山　直毅（北九州市立大学）

米国大統領（以下，大統領）は毎年１月に一般教書演説（State of Union Address）を行い，その時の関心事を連邦議会に対し演説を行う。本研究ではこの演説原稿をコーパスとし，トピックモデルと呼ばれる手法を用いて大統領の主要政治課題がどのような社会背景に影響を受けてきたのかを明らかにする。

本研究では，様々なウェブサイトで公開されている一般教書演説を１つにまとめ，表１のようなコーパスを作成した。このデータに対し，ストップワード（田畑（2017）が利用したものを改訂）処理を施し，潜在的ディリクレ配分法（Latent Dirichlet Allocation）（Blei 他，2003）によってトピックを解析することで，35個のトピックから３個の主要トピックを得ることができた。

表１：一般教書コーパス情報

総ファイル数	トークン頻度	タイプ頻度
228	1,751,570	32,204

まず，米国の初代大統領が就任した1790年から1900年までの110年間と，その後，断続的に1916年までの間，*congress* や *government*，*country*，*duty*，*duties*，*law(s)* といった語彙が多く現れる。これらのうち，*duty* のコロケーションを調査すると，大統領の仕事を強調する *my duty* という表現は134例のうち116例がこの期間に現れている。また *duties of the federal/general government* といった，政府の役割を強調する表現は27件中26件がこの期間に現れる。このことから本トピックは「国の役割」と言える。

次に，1914年，1915年，1917年，そして1932年から1980年までの間，*peace* や *world*，*freedom*，*national*，*security*，*defense* といった表現が頻出するようになる。これらの語彙のコロケーションを見ると，*peace and freedom in the world* や *national security*，*national defense* といった表現が目立つ。そのため，この時代の主要トピックは「軍事」であったと言える。

最後に，1960年代からトピックの重要性が上がり始め，1982年以降，最も重要であるとされるトピックは，*jobs*，*children*，*families*，*health*，*care*，*working* といった語彙が目立つ。これらをコンコーダンスラインで確認すると，*working families* や *health*

insurance, *create new/good/more jobs* といった表現が目立つ。このことから，近現代の大統領の関心事は「社会福祉」に関してであると言える。

以上を総括すると，大統領の関心事は［[国の役割 => 軍事関連 => 社会福祉]］というトピックの変化を辿ってきたことになる。本研究では，一般教書演説のトピックは建国時の内政，世界大戦や冷戦，そして冷戦後の米国内景気の悪化という社会背景に影響されていることを，コロケーションなどの観点から論じる。

■９月30日（土）

【研究発表第５セッション】

【研究発表１】

コーパスの示す科学的データと学習性・商品性との両立
―『ウィズダム英和辞典』の編集にあたって ―

井上　永幸（広島大学）・西垣　浩二（㈱三省堂辞書出版部）

『ウィズダム英和辞典』は，本格的にコーパスを活用して編纂された初の英和辞典として，初版以降2回の改訂を行い，現在第３版が刊行されている。コーパス分析の成果をいかに盛り込んでいるかというような点については，過去に本学会やその他の機会に何度か発表をしているが，本発表ではコーパスから得たデータ，あるいはそこから漏れるものを，いかにして掘り下げ，学習者にとって有益な記述を作り上げてゆくかという点について，実例をもとに論じることとしたい。

例えば「頻度」や「コロケーション」などについて，コーパスの分析により得られる知見は，コーパスを用いた辞書編纂にあたってはもっとも参考とすべきところであるが，当然「学習者が効率的に学習を進めるための教材」を編纂するという立場に立つのであれば，頻度の高い表現であれば何でもすべてを採用するという考えには立てない。学習性，日常生活への密着度（これはコーパスには反映されにくい場合もある），記述分量と学習効果のバランスなど，さまざまな点について熟慮を加えたうえで，紙面に反映しているのである。これは，語法･類義解説，用例採用基準，語義分析など，辞書のすべての記述において言えることである。コーパスの産出する科学的データはそれ自体が目的となるのではなく，それをもとに学習者にとって必要な情報を取り出し，いかに学習効率が上がるように情報を配置してゆくかという点が，学習者にとって使いやすく，結果的に商業的にも成功する教材を編纂するにあたって肝要な点である。

本発表においては，コーパスから得たデータをいかに昇華させ，学習効率の高い教材を編纂してゆくかという点について論じることとしたい。また，開示できる範囲内で，コーパス構築・活用の裏話等にも言及する。井上からは，『ウィズダム英和辞典』における記述について，実例を挙げながらどのようなデータを反映している

のか，コーパスデータに「加えて」，どのような点を編者の「経験と勘」によって補っているのかについて論じる。西垣からは，成功する「商品」としての辞書の編集にあたって，記述内容とターゲットユーザの求める情報との整合性，記述分量のバランスなどについて補足をする。

【研究発表２】

英語辞書レーベルとコーパス

田畑　圭介（神戸親和女子大学）

現在の英語学習辞典において学習者に語彙情報を効果的に伝達する手法の一つにレーベルの提示がある。レーベルは該当語の用法だけでなく他の語との差別化を図る機能も持ち合わせている。本発表では最初にレーベルにヘッジが付与されうる集合タイプのものと付与されない段階的タイプのものに二分されることを例証する。そして［disapproving］と［offensive］のレーベル，［informal］と［spoken］のレーベルへの細分化の必要性をCOCAとテレビドラマコーパスのデータをもとに論証する。Longman Dictionary of Contemporary English（LDOCE）では［offensive］は採用しておらず，Oxford Advanced Learner’s Dictionaryでは［spoken］は採用していないが，condescending，conventional，brigade，bowdlerize，foolhardyといった語の使用状況から，［disapproving］，［非難して］のレーベルの必要性，foreigner，fruitなどの使用状況から［offensive］，［けなして］のレーベルの必要性が帰結される。you could have fooled meはCOCAでFIC10例，SPOK ３例，MAG １例，NEWS １例の計15例検出されるが，各用例はいずれも会話文で用いられている。これは会話で用いられる表現であることを示すと共にCOCAのジャンル分析に注意が必要であることも暗示している。fortunatelyとluckilyをCOCAのSPOKENで検索すると，fortunately 1107，luckily 532となる。日常的な表現であるluckilyの使用が予想外に少ないが，これらはCOCAのSPOKENがくだけた会話体でなく報道番組の発話データが中心となっていることに起因している。これらの事実はfortunatelyを［ややかたく］としてluckilyと差別化し，LDOCEのようにspoken ≠ informalと捉える必要性を示すものとなる。

【研究発表３】

怒りを表す類義語と概念メタファー

南澤　佑樹（大阪大学大学院生）

本発表の目的は，コロケーションを抽出する統計手法を用いてメタファー表現を収集し，怒りを表す類義語 *anger*，*rage* に見られる概念メタファーの違いを示すことである。

感情のメタファーに関しては，概念メタファー理論の枠組みに基づき盛んに議論が行われてきた。怒りの感情に対しても複数のメタファーが提案され，怒りを容器内の熱い液体とみなす ANGER IS A HOT FLUID IN A CONTAINER（“*boiling with anger*”）はその中でも最も中心的なメタファーとされている（Kövecses, 1990, 2000）。しかしながら，多くの先行研究では *anger* と *rage* が区別されておらず，類義語間の違いにはこれまであまり関心が向けられてこなかった。この理由には，従来のアプローチでは結果を量的に分析することが困難であったという点が挙げられる。

近年では，概念メタファー分析にもコーパスが用いられるようになっているが，現段階ではメタファー表現をコーパスより網羅的に収集することは難しい。これを踏まえ Stefanowitsch（2006）は，根源領域に属する語と目標領域に属する語（本発表では感情語）が共起する表現であるメタファーパタンを分析する手法を提案している。Turkkila（2014）は，この手法を用いて怒りの類義語に見られる概念メタファーを分析しており，*anger, rage, fury* に見られるメタファーが概ね同じであると主張している。しかし，コーパスを用いた手法では，メタファー表現の出現頻度によって概念メタファーの重要度を決定することが多いことから，*boiling with anger* や *his anger welled up* のような感情の様態を具体的に表す表現よりも *in anger* のような意味内容の薄いメタファー表現を重要とみなす傾向がある。しかしながら，そのようなメタファー表現は他の感情等にも用いられる一般性が高い表現であるため，それらによって類義語間の違いを見いだすことは難しい。

これらを踏まえ本発表では，British National Corpus よりそれぞれ *anger, rage* を検索語としてメタファー表現を収集し，その相違点を指摘する。本研究では，抽象的で一般性の高いメタファー表現ではなく怒りの様態を具体的に表すメタファー表現を収集するため, 意味的に結びつきの強いコロケーションを測る MI スコアを用いる。その結果，*anger* は *vent, seethe, well, simmering* といった語と結びつきが強く，従来の主張通り ANGER IS A HOT FLUID IN A CONTAINER が最も中心的な概念メタファーであると言えるのに対し，*rage* では，*anger* と比較して *howl, bristle, murderous* といった語がリストの上位にきていることから ANGER IS A HOT FLUID IN A CONTAINER だけでなく ANGER IS A DANGEROUS ANIMAL（“*bristling with rage*”）とも結びつきが強いと言えることを主張する。また *anger* と結びつくメタファー表現は怒りの様々な側面を表すのに対し，*rage* と結びつきの強いメタファー表現は怒りの特定の側面

を表す傾向にあることも主張する。

■９月30日（土）
【研究発表第６セッション】
【研究発表１】

日英対訳コーパス中の「～ことになる」構文とその英訳文間の構造的不一致

大矢　政徳（目白大学）

機械翻訳の分野における Bitext word alignment（BWA）手法では，翻訳元の文中の単語と翻訳先の文中の単語との対応関係（単語アライメント）を統計的に算出する。しかしながら，日本語の表現には，英語では複数の表現で翻訳される場合や，日本語には存在しない要素を補完しなければならない場合が高頻度で存在し，この領域は BWA 手法ではカバーしきれない。例えば，「健は明日帰国することになっている」という日本語文は，英語では典型的には ”Ken is supposed to come back to Japan.” と訳され，「～ことになる」は英語では ”be supposed to” に対応しているが，実際の「～ことになっている」構文は，必ずしも ”be supposed to” と訳されているとは限らない。また，日本語では主語や目的語が省略される場合が多く，それらを英語のように主語や目的語が省略される頻度が低い言語へと翻訳する場合に問題となる。

本研究では，このように BWA 手法ではカバーしきれていない領域を補完することを目的とし，翻訳元文と翻訳先文との対応関係を，単語間のアライメントだけではなく，単語間依存関係木の統語的不一致パターン（syntactic divergence patterns）として提示することを提案する。コーパスデータとして『Wikipedia 日英京都関連文書対訳コーパス』の日本語文と英語対訳文をそれぞれ構文解析して得られた単語間依存関係情報を用い，日本語構文「～ことになる」を含む日本語文1197文から無作為に100文を選び，これらの文中の「～ことになる」が英語対訳文のどの単語・フレーズに対応しているかを統語的不一致パターンとして人手で抽出し，各パターンの発生確率を算出する。特に，依存文法の枠組みでは，統語依存木の根の位置にある述語（主節の動詞がこれに該当する場合が多い）がどのような要素を述語項として要求するか（例：形容詞を含む名詞句か，それとも関係節を含む名詞句か）がその統語依存木全体の構造を決定する点を鑑み，日本語構文「～ことになる」が統語依存木の根の位置にある場合に，これに対応する英語対訳文では統語依存木の根の位置にないという統語的不一致パターンの発生確率が高いことを示す。さらに，統語的不一致パターンを人手によらず自動で抽出することを目的として，当該100文から得られた「～ことになる」統語的不一致パターンを，上述コーパス中の「～ことになる」を含む日本語文1197文からこれら100文を除いた1097文からさらに無作為に選んだ100文に対して，正規表現でマッチングすることによって自動的に抽出する手法を試み，

その可能性と改善点について論じる。

【研究発表２】

医学研究論文ジャンルにおけるコーパス作成ツール AntCorGen を活用した教育の可能性
—Construction of Corpora for Discipline-Specific Learning in Medical Research Article Genres

浅野　元子（大阪大学大学院生）

本発表の目的は，最近公表されたコーパス作成ツール AntCorGen（Anthony, 2017a）を用いて構築した比較的大規模な医学研究論文コーパスにおけるテクストの言語的特徴を量的質的に検討し，教育応用の可能性を探索することである。

英語が医学論文での国際共通語となり，小規模言語を母語とする大学院生や研究者（Giannoni, 2008）は専門家集団に受容されるために分野特有の修辞パターンの習得が必要といわれる（Flowerdew, 2013）。研究論文を用いてミニコーパスを構築するデータ駆動型学習（Data-Driven Learning）が提案されて久しいが（Anthony, 2017b; Lee & Swales, 2006; Noguchi, 2004; 朝尾・投野，2005），多様な専門分野の範囲をどのように限定して論文を使用すべきかについての報告は少ない。本稿では，医学研究論文の言語的特徴を検討した研究（Nwogu, 1997）を参考に，同一誌での種々の医学研究論文に異なる言語的特徴があるかどうかについて AntCorGen に実装された PLOS ONE 誌の学術論文をコーパス化する機能を用いて検討した。

医学･健康科学分野のうち Cardiology（心臓病学）, Gastroenterology and hepatology（胃腸病学と肝臓学），Pulmonology（呼吸器学），Oncology（腫瘍学）の領域での論文各5000報を得た後に100報ずつを無作為抽出した。個々の論文をタイトルならびに抄録におけるムーブ（Swales, 1990; Salager-Meyer, 1990 & 1992）とヒント表現（Tojo, Hayashi, & Noguchi, 2014）を手がかりに医学研究の種類すなわち動物での研究や症例コホート研究など（国立国際医療センター，2009）に分類した。論文テクストの語彙を計量し（Imao, 2015），本文の頻度上位語を変数としてウォード法，ユークリッド距離を用いてクラスター分析を行った（田畑，2004）。

本文は総語数が1,727,472語，異なり語数を総語数の平方根で除して算出した Guiraud Index が36.8で，内容が詰まった文書であることが示唆された。語彙によるクラスター分析では医学研究の種類による類似性が示唆された。

本ツールは学術論文コーパス構築において利便性が高く有用であると考えられた。教育現場では，本ツールを用いて構築したコーパスにおける論文からタイトルや抄録を頼りに対象研究と同一種類の研究に関する論文を選択して各自のコーパスを作成すると，目標とする専門家集団が慣れ親しむ修辞パターンをより実践的に学習することが可能性となることが示唆された。

【研究発表３】

Applying Topic Models to Describe a Corpus's Compositionality: How can the external criteria be associated with meaningful sets of internal evidence?

Tomoji Tabata（University of Osaka）

Topic modelling is a machine learning method for uncovering hidden semantic structures in a corpus of texts. Based on a probabilistic inference algorithm, latent Dirichlet allocation, the technique makes it possible to identify sets of frequently co-occurring words, or topics, that characterize a text as well as classify texts into meaningful groups defined by inferred sets of strongly associated topics.

One of the major advantages topic modelling has over traditional key-word detection techniques, such as the Chi-squared test, log-likelihood ratio test, Mann-Whitney's *U* (or exact rank) test, or somewhat modestly but robustly applied Welch's *t*-test, employed in many stylometric/corpus linguistic studies is that topic models do not simply provide typical dichotomous or polarized sets of key-words for a target corpus versus a reference corpus, but enable us to spotlight key-words of multiple sets of texts, thereby making it possible to classify texts into reasonable subsets clustered in terms of word co-occurrence patterning. Outputs obtained from a topic modelling run range from a word-topic association table, which tells us what and how many topics a particular word belongs to and how much weight the word has in each topic; a topic composition table, which illustrates what types of words knit up a particular topic and to what extent individual words contribute to composing a given topic; to a text composition table, which accounts for topic density in each of analyzed texts, or to what extent a text is occupied by words belonging to each of the topics associated with the text. Of further interest in the context of corpus linguistics is that results of a topic modelling can be visualized in the form of a topic box-plot, network diagram of topics and words as well as that of topics and texts, and a summarizing heatmap of topics and texts under investigation.

The present study applies topic modelling to the FLOB corpus with a view to analyzing latent semantic structures/patterns underlying in the corpus and mapping its subcorpora (or, registers) in the network of words, topics, and texts. What is of special interest is that by means of this approach it is now possible to shed new light on thematic/topical structures composed by a large number of infrequent words, which would otherwise escape the net of key-word statistics due to infrequency of occurrence or lack of a proper classification of lexical items.

This paper summarizes results of multiple topic modelling runs on the FLOB corpus. The paper reports that the text categories A—J (informative prose registers) are clearly distinguished from the texts that belong to the imaginative prose writings, or fiction (categories K—R) according to topical structures underlying in the corpus. The fifteen text registers in the

FLOB corpus are classified into the two distinct clusters: informative versus imaginative proses. To turn our attention to topic distributions across registers, we can notice that the generated topics are divided into two sets: topics contributing more to the informative registers and those talked about more in the fictional registers.

Emerging results from this research are expected to open up a new avenue of inquiry into key semantic patterns in a large collection of texts, thereby suggesting a possibility of building a bridge between findings from machine learning text mining and traditional stylistics, distant reading and close reading, with an empirical interplay of insights that will benefit modern text analysis. Of further note is that in interpreting results of a topic modeling, we are likely to confront a sea of multitudinous contentious interpretations. Topic modeling is not just a cutting-edge machine-learning technique, it involves a highly humanistic interpretation and insight: the true value of machine-learning can only be judged by the depth of human insight, ironically but interestingly.

■10月１日（日）
【ワークショップ２】

機械学習を用いたコーパス分析入門

小林雄一郎（日本大学）

本ワークショップでは，近年コーパス言語学の分野でも盛んに利用されるようになってきた機械学習（machine learning）の技術を紹介します。機械学習は，人間が持つ学習能力をコンピュータに持たせることを目指す人工知能の研究分野です。また，コンピュータにデータを解析させることで，データの背後に潜むパターンを発見（学習）させる技術のことを指します。そして，多くの場合，データから発見されたパターンは，新たなデータの予測に活用されます。

機械学習の技術を用いることで，手作業では扱えないような大量のテキストデータを効率的に分析できるようになります。そして，パターンを発見するための十分な量のデータを用意すれば，人間が予測するよりも高い精度で予測を行うことが可能になります。さらに，予測に寄与したパターンを吟味することで，分析対象のテキストを特徴づける言語項目を特定することができます。

コーパス言語学における機械学習の活用事例としては，テキストの著者推定やジャンル推定，英作文の自動採点，語彙や文法の使用に関する通時的分析などがあります。本ワークショップでは，このような事例を紹介しつつ，機械学習の基本を講義形式で詳しく説明します（ハンズオンの実習形式ではありません）。

ワークショップの流れとしては，(1) 機械学習とは何か，(2) データの準備方法，(3) 具体的な仕組みと手順，(4) 分析結果の検証方法，(5) コーパス言語学における活用

事例，を予定しています（諸般の事情で若干変更する場合があります）。なお，本ワークショップは初学者を対象としており，統計学などの事前知識を参加者に求めません。また，機械学習の手法を説明するにあたっては，可能な限り，分かりやすい言葉やイメージを使うことを心がけ，四則演算（足し算・引き算・掛け算・割り算）以外を使った数式は出しません。

【講演】

A Frontier in Learner Corpus Studies: For Better Understanding of L2 Learners

Shin’ichiro Ishikawa（Kobe University）

Various learner corpora have been developed to date and they have greatly contributed to improvement of L2 teaching. However, a more carefully designed corpus would be needed for a reliable contrastive interlanguage analysis. Thus, recent learner corpora have come to pay much more attention to controlling variety in the collected data.

The International Corpus Network of Asian Learners of English (ICNALE) is one of the largest learner corpora ever compiled. It includes more than 10,000 speeches and essays produced by L2 English learners in ten countries and regions in Asia as well as English native speakers. Its unique feature is that the topics are carefully controlled. All the participants are required to speak or write about two kinds of common topics: (A) It is important for college students to have a part-time job and (B) Smoking should be completely banned at all the restaurants in the country. Such a topic control is expected to lead to a greater reliability in varied types of contrastive analyses (Ishikawa, 2013).

The ICNALE currently consists of four modules: Spoken Monologue (1,100 participants, 4,400 samples, 500,000 tokens), Spoken Dialogue (under construction), Written Essays (2,800 participants, 5,600 samples, 1,300,000 tokens), and Edited Essays (290 participants, 580 samples, 140,000 tokens).

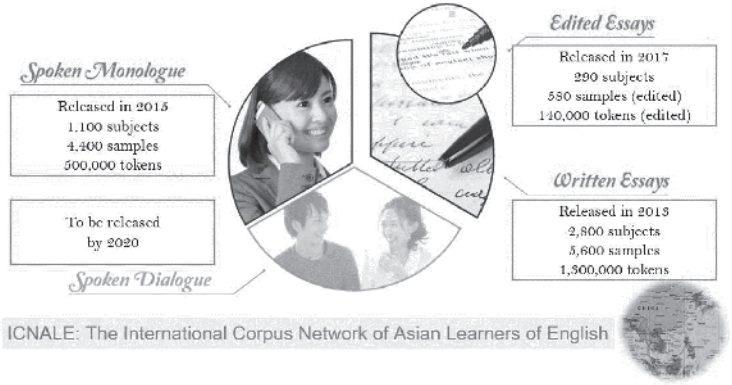


Fig. 1 The structure of the ICNALE

The ICNALE development team believes that comparing something comparable is a key to further development of learner corpus studies.

【シンポジウム】

話し言葉コーパスの構築と利用

司会：野口ジュディー（神戸学院大学名誉教授）

コーパス言語学の多くの研究は書き言葉を扱っていますが，人間の自然言語は元々話し言葉からスタートしています。しかし，話し言葉は扱いにくいものです。コーパス作成には，発話者から許可を得ないとレコーディングすることさえ難しいでしょう。また, 生の言葉を捉える環境がレコーディングに適していません。このようなハードルを越えての，時間のかかる書き起こしが必要になります。最終的にはコーパスを利用しやすいインタフェースを用意しなければなりません。そういった作業の苦労話を研究者たちは少なくともいくつかは抱えています。そういった研究者たちの努力によって，異なる４つコーパスが利用できるようになりました。英語を lingua franca として使用する，様々な母語を持つ学習者（大学生や大学院生を含む）からプロフェショナルまでの話し言葉がそれにあたります。この４つの異なるコーパスをどのように研究や教育に利用できるかを紹介いたします。各講師が構築したコーパス〔学習者の書き言葉・話し言葉（英語)〕ICNALE，学習者話し言葉日本語（テーマ別）I-JAS，話し言葉日英（理系プレゼン）JECPRESE，TED コーパス〕に関して，構築とその利用方法について話していただきます。

The ICNALE：中間言語対照分析の精緻化とアジアにおける学習者コーパス研究の発展を目指して

石川慎一郎（神戸大学）

The ICNALE（The International Corpus Network of Asian Learners of English）は，アジア圏10か国・地域において，英語学習者の L2 産出データを収集するプロジェクトで，すでに，Written Essays，Spoken Monologue，Edited Essays の３つのモジュール（計約190万語）が公開され，現在は，Edited Essays モジュールの拡充と，初のマルチモーダル版となる Spoken Dialogue モジュールの開発が進められています。The ICNALE の特徴は，プロンプトやテキストの長さなどが一定の範囲で統制されていることで，これにより，信頼性の高い国際比較研究が可能となります。The ICNALE は，ダウンロード版のほか，オンライン版があり，専用の検索用インタフェースが開発されています。

International corpus of Japanese as a second language: 日本語学習者の言語研究と指導のために

迫田久美子（広島大学・国立国語研究所）

International corpus of Japanese as a second language（I-JAS，http://lsaj.ninjal.ac.jp/ ）は，12の言語を母語とする日本語学習者の発話と作文のコーパスです。JFL 学習者と JSL 学習者，さらに国内の学習者は教室環境と自然環境学習者のデータが収められ，さらに同じタスクを実施した日本語母語話者のデータも含まれています。完成は2020年春を予定しており，現在は450名のデータが公開されています。I-JAS の特徴としては，英語，中国語，韓国語，西語，独語，仏語，露語，タイ語，トルコ語，インドネシア語，ベトナム語，ハンガリー語の母語話者の複数のタスクのデータを所収しており，検索システムを備えていることが挙げられます。全員が同じテストを受けており，成績や背景事情も公開しています。

JECPRESE: JSL と EFL ユーザーのために

野口ジュディー（神戸学院大学名誉教授）

JECPRESE, the Japanese-English Corpus of Presentations in Science and Engineering（http://www. jecprese.sci.waseda.ac.jp/），は留学生のための専門日本語教育（JSL, Japanese as a second language）を支援する研究発表コーパスでスタートしました。日

本の大学院生の日本語プレゼンテーションに加えて，アメリカの大学生や国際学会の英語プレゼンテーションも収められていて，EFL（English as a Foreign Language）の学生にも利用できるコーパスになりました。理工系のプレゼンテーションの特徴をわかりやすくするために，各発表を ESP（English for Specific Purposes）の手法であるジャンル分析に基づいてセクションやステップで検索できるようにしました。単語や表現の検索もできます。

TED Corpus Search Engine: TED Talks を研究と教育に活用するためのプラットフォーム

長谷部陽一郎（同志社大学）

TED Corpus Search Engine（http://yohasebe.com/tcse）は TED Talks の英語トランスクリプトを検索するためのウェブ・システムです。定期的な更新によってデータは増え続けていますが，現時点では約2400のプレゼンテーションから抽出されたテキスト（延べ語数は約600万語，異なり語数は約８万語）が収められています。語の品詞や基本形を指定可能な検索機能を実装しており，特定の構文や談話標識の実例を取得して研究や教育に役立てることができます。その他の特徴としては，得られた発話セグメントの動画をピンポイントで再生する機能や，日本語を含む28の言語による対訳データを表示／検索する機能が挙げられます。本発表では本システムを言語研究に活用する方法を中心に論じていきたいと思います。

『英語コーパス研究』投稿規定

（2017年９月29日改定）

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70字×35行（Microsoft Word の場合は全角の文字数を指定するため，35字を指定する），周囲の余白１インチ（25.4mm），6,000語以内（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字×30行, 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 を原則とするが，執筆者の意向でフォントの種類を変えてもよい。

（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70字×35行（Microsoft Word の場合は全角の文字数を指定するため，35字を指定する），周囲の余白１インチ（25.4mm），4,500語以内（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字×30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も Times New Roman を原則とするが，執筆者の意向でフォントの種類を変えてもよい。

（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

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。

第２条　本会はコーパスを用いた英語およびその関連領域の研究を促進することを目的とする。

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．大会の開催（年１回）
２．会誌・会報の発行
３．その他本会の趣旨に沿う事業

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，およ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，大学また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．名誉会員は本会の活動に特別に寄与したものとする。

第５条　本会の会員は所定の会費を納めるものとする。会費の額については別にこれを定める。

第６条　本会に次の役員をおく。

１．会長　　１名
２．副会長　　１名
３．事務局長　　１名
４．理事　　若干名
５．監事　　１名
６．会計　　１名
７．顧問　　若干名

第７条　会長は，理事の互選によって選出する。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　会長は本会を代表し，会務を統括する。会長は総会・理事会を招集し，これを主宰する。

第８条　副会長は, 理事会の承認を受け会長が委嘱する。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　副会長は会長を補佐し，必要に応じて会長の職務を代行する。

第９条　事務局長は，理事会の承認を受け会長が委嘱する。事務局長の任期は２年とし，引き続き２期までの再任を妨げない。

2　事務局長は本会の運営に関わる諸事務を担当する。
3　事務局長は会務を円滑に遂行するため，理事会の承認を受け事務局補佐を置くことができる。
第10条　理事の委嘱は，理事会の承認を受け会長が行う。理事の任期は２年とし，引き続き再任を妨げない。
2　理事は理事会を組織し，本会の運営に関わる重要事項を審議する。
第11条　監事は，理事会の承認を受け会長が委嘱する。監事の任期は２年とし，引き続き２期までの再任を妨げない。
2　監事は本会の財産および会計の状況を監査する。
第12条　会計は，理事会の承認を受け会長が委嘱する。会計の任期は２年とし，引き続き２期までの再任を妨げない。
2　会計は本会の会計全般を担当する。
第13条　本会に顧問をおくことができる。顧問は役員を退任したもので，本会の発展に特に功績のあったものとする。
第14条　本会は毎年１回総会を開く。総会は会則の改定，予算・決算その他重要事項を審議する。
第15条　本会に次の委員会を置く。各委員会の規程は別に定める。
１．編集委員会
２．学会賞選考委員会
３．大会企画委員会
第16条　本会に研究会を置く。研究会の規程は別にこれを定める。
第17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。
付則　この会則は2000年４月１日から施行する。
付則　この会則は2009年４月１日から施行する。
付則　この会則は2010年４月１日から施行する。
付則　この会則は2011年４月１日から施行する。
付則　この会則は2014年４月１日から施行する。
付則　この会則は2015年４月１日から施行する。
付則　この会則は2017年４月１日から施行する。
細則（第５条）本会の会費は次の通りとする。
１．一般会員年額 6,000円（在外会員は年額12,000円）
２．学生会員年額 3,000円（在外会員は年額10,000円）
３．団体会員年額 5,000円
４．賛助会員年額15,000円
５．理事年額10,000円
６．名誉会員ならびに顧問からは会費を徴収しない。
７．会費は年度始めに納入するものとする。また，原則として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。

英語コーパス研究（第25号）
【2018年3月31日発行】

編集・発行　ⓒ2018 英語コーパス学会
〒157-8511 東京都世田谷区成城6-1-20
成城大学 社会イノベーション学部 石井康毅研究室気付
E-mail（事務局長）: jaecs.hq@gmail.com
Twitter: @JAECS2012
Website: http://jaecs.com/
郵便振替口座：00930-3-195373（英語コーパス学会）
印　刷　所　株式会社ニシキプリント
〒739-2117 東広島市高屋台2丁目1番12号

English Corpus Studies: Vol.25
2018

JAPAN ASSOCIATION FOR ENGLISH CORPUS STUDIES