A corpus-based investigation of collexemes for active-passive alternation in the English part of an English-Japanese parallel corpus

Masanori Oya
Meiji University
masanori_oya2019@meiji.ac.jp

Abstract

This study conducted a corpus-based investigation of collexemes for active-passive alternation found in the English part of an English-Japanese parallel corpus as an attempt to use them as metrics for distinguishing native English and non-native English. The results show that some verbs in the data are used more often in the active voice than the passive voice, and vice versa, and the differences are statistically significant. However, these verbs are not the same as those found in a previous study. This fact supports the claim that active-passive alternation constitutes a lexico-semantic phenomenon that is sensitive to various factors, such as differences in genres and type of the authors of the text.

1. Introduction

This study conducts a corpus-based investigation of *collexemes* (Gries & Stefanowitsch, 2004; Stefanowitsch & Gries, 2003) for active-passive alternation found in the English part of an English-Japanese parallel corpus. Collexemes are a set of words that are attracted to certain types of *constructions* in the sense of the term used in Goldberg (1995) and Lakoff (1987). Collexemes are used in a certain construction more often than other words, and the difference between the frequency of their use and that of non-collexemes in the same construction is statistically significant. Investigation of collexemes in constructions is expected to facilitate a deeper understanding of the relationship between the lexicon and syntax. This is because such investigations allow us to have a critical viewpoint about the mainstream syntactic investigations of today, which presuppose that words are inserted into certain syntactic structures arbitrarily without considering other factors such as semantics of the word and discourse or genre of the text wherein the sentence is used. Rather, an investigation of collexemes is expected to suggest that the lexicon and syntax are closely related to each other, and different syntactic constructions necessarily attract certain words because of their semantic properties and other factors dependent on the characteristics of each construction.

This study is an attempt to use collexemes as one of the metrics for distinguishing native English and non-native English. If a collexeme in English texts generated by native speakers of English is found as non-collexeme in English texts generated by non-native speakers of English, then that collexeme indicates the difference between these two groups of speakers of English. In addition to this, the information of collexemes is expected to be of educational value; learners can learn different collexemes for different constructions and that will lead them to more natural use of the language.

The remainder of this paper is structured as follows: Section 2 reviews the rationale of collexeme investigations in contrast with collocational analyses. Section 3 reviews previous studies on collexemes, with special attention on the works of Gries and Stefanowitsch. Section 4 describes the data used in this study. The method, results, and discussion on the results are reported in Sections 5, 6, and 7, respectively. Lastly, Section 8 concludes.

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 516-521, Hakodate, Japan, September 13-15, 2019

2. Collocation and Collexemes

One important aspect of corpus linguistics is *collocational* analysis, in which the semantic and syntactic properties of a word or phrase are analyzed in terms of the context in which the word or phrase appears. Context here refers to the words before and after the word or phrase to be investigated, and they are called *collocates*. The span of collocates varies across different researchers with different research interests; for example, it is ±1 in Kennedy's study of *between* and *through* and ±5 in Church and Hunk's analysis of *doctor* (as cited in Stefanowitsch & Gries, 2004).

The problem with collocational analysis is that it only focuses on the linear order of the target word or phrase and its collocates, and it ignores their syntactic relationships. In this respect, Stefanowitsch and Gries (2003) point out that collocational analysis cannot capture the deeper relationship between the target word or phrase and the words associated with it, or the relationship between the target word and certain *construction pairs* that are considered as alternates. The so-called key word in context (KWIC) cannot capture the difference between construction pairs. For example, it cannot capture the context of where the target word appears in the double-object construction and where it appears in the direct-object-to-object construction (e.g., Sarah has given David some books vs. Sarah has given some books to David). In linguistics, these are called *dative alternations*, and it is virtually impossible to capture such alternations in KWIC, because they appear across words in such sentences, and the linear order of these words does not contain enough information to represent each alternation.

Stefanowitsch and Gries (2003) first introduced the idea of *collostructure* and applied it to corpus data, in order to overcome the shortcomings of KWIC-style research stated above. Their research aimed to apply the idea of *construction* (Goldberg, 1995; Lakoff, 1987) into the investigation of significant associative relationships between vocabulary and grammatical structure. They assumed that 1) lexicon and grammar are fundamentally similar and 2) multi-word expressions create links between the lexicon and grammar. Their assumptions can be paraphrased as follows: the so-called alternating constructions such as *active-passive alternations* (e.g., Sarah has broken some dishes vs. Some dishes have been broken by Sarah) and dative alternations do not alternate from one construction to the other, but they are actually two different constructions that are independent from each other. Stefanowitsch and Gries argue that this means these construction pairs should not be treated syntactically but lexico-semantically; some words are attracted to one of the construction pair, while others to the other. In other words, certain sets of verbs are used more often in the active voice, while another set of verbs are used more often in the passive voice, and the difference in their frequencies is statistically significant. Collexemes are such words that are attracted to certain constructions.

Inspired by their investigations, this study explores the possibility that English texts with limited focus on a topic contain certain collexemes for certain constructions. In particular, this study conducts a corpus-based investigation of collexemes for active-passive alternations found in the English part of an English-Japanese parallel corpus, which was constructed by translating a Japanese original text into English (the details are described in Section 4). These data are selected because it is expected that the collexemes for active-passive alternations reflect the characteristics of non-native English in terms of the collexemes for the alternation, which are different from the collexemes found through research using the corpus data generated by native speakers of English. As mentioned in the previous section, this study constitutes an attempt to use collexemes as one of the metrics for distinguishing native English and non-native English, with their educational value in mind.

3. Previous Studies

Stefanowitsch and Gries (2003) investigated constructions such as *cause N* (nouns that are attracted to the verb *cause*), X think nothing of V *gerund*, into-causative (e.g., Sarah tricked David into employing her), ditransitives, progressives, the imperatives, and past tense, and they found collexemes for each of these constructions. Based on the same assumption, Gries and Stefanowitsch (2004; G&S henceforth) conducted further research on the constructions' *active-passive alternations* and future tense as *will* and *be going*

to. They found that each construction is associated with a set of collexemes, and the association is so strong that there is a statistically significant difference between the frequency of these verbs in the active voice and the passive voice. The same is true for the pair of *will* and *going to*.

Investigations of collexemes are extended to the study of *semantic prosody*. Semantic prosody is a phenomenon where a certain word is associated with a positive or negative connotation because of its frequent occurrence with certain other words (Sinclair 1991). For example, Tang (2017) showed that the verb *cause* appears in various constructions which also contain words with negative semantic connotations, and therefore the verb *cause* has the tendency to be accompanied with negative semantic prosody.

4. Data

The corpus used in this study is the Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles, v.2.01 (National Institute of Information and Communications Technology, 2011). This corpus includes approximately 500,000 Japanese-English translation pairs of Wikipedia articles on 15 topics related to Kyoto, and each topic comprises one subcorpus. The articles are translated from the original Japanese text into English manually by Japanese translators, and then these are proofread by native English speakers. These translations are then edited by Japanese professionals, with special attention paid to the technical terms. This study uses the subcorpus of the topic related to Buddhism, which contains 26,890 Japanese-English translation pairs. These data are chosen with the assumption that English sentences translated from Japanese sentences are one of the genres of non-native English.

5. Method

In this study, the English sentences in the data are parsed by the Stanford Dependency Parser (de Marneffe & Manning, 2008), and the parsed results are used to calculate the number of verbs associated with their subject and object (active transitive verbs) and with their subject in the passive voice (passivized transitive verbs). We can count the number of passive verbs in the corpus by counting the number of dependency-type nominal subject of passivized verbs (NSUBJPASS in the parsed output) in the parse output through a simple regular-expression search. As for active verbs, on the other hand, we can determine their number by counting the dependency-type direct objects (DOBJ in the parsed output) in the same parse output through the same search method as for passive verbs. This means that this study ignores active verbs that are used without their direct objects with the assumption that they are used as intransitive verbs and therefore should not be counted as transitive verbs.

To show that the difference is larger than a coincidence between the probability that a verb *v* is used in the active voice in the corpus data and that all the verbs other than *v* are used in the active voice in the same corpus data, we conduct Fisher's exact test (1922, 1954), which was developed to examine the significance of the association between the two groups. This test has the following characteristics: it can be used when 1) the sample size is small and 2) the data are not distributed normally. This study uses this test because of these characteristics, as was the case in G&S.

In addition, to show exactly how large the difference between these two probabilities is, we calculate Cohen's *h* (Cohen 2013), which G&S did not. Cohen's *h* is employed to measure the differences between proportions in relation to hypothesis testing. The difference between two proportions is "statistically significant" when it seems that the population proportions are different. However, it is also possible that this difference can be too small to be meaningful. In other words, the "statistically significant" result does not indicate how large the size of the difference is. In this context, Cohen's *h* indicates the size of the difference and allows us to decide how meaningful the difference is.

Cohen's *h* is calculated in the following procedure. First, each probability is transformed through an "arcsine transformation" as follows:

$$\varphi = 2\arcsin\sqrt{p} \qquad (1)$$

When we have two probabilities, *p1* and *p2*, Cohen's *h* is the difference between their arcsine transformations:

$$h = \varphi 1 - \varphi 2 \qquad (2)$$

Cohen's *h* is interpreted as follows through a rule of thumb:

h = 0.20, "small effect size"; *h* = 0.50, "medium effect size"; *h* = 0.80, "large effect size."

In this study, $\varphi 1$ is the probability that a given verb *v* is used in the active voice, and $\varphi 2$ is the probability that all the verbs other than *v* is used in the active voice. For each verb in the data of this study, Fisher's exact test and Cohen's *h* are calculated by using js-STAR ver. 9.2.5j (http://www.kisnet.or.jp/nappa/software/star/freq/2x2.htm#). If Cohen's *h* is larger than 0.8 for a verb, the verb is more likely to be used in the active voice, while if it is smaller than -0.8, the verb is more likely to be used in the passive mood. If Cohen's *h* is between -0.2 and 0.2 for a verb, the verb has no preference of being used either in the active or passive voice. We ignored such verbs that appear less than 15 times in the data, either in the active or passive voice, so we can concentrate on frequently used verbs.

6. Results

This study found that the data contain 4,751 active verbs and 2,765 passive verbs. These total 7,516 verbs belong to 960 types, of which 306 are used in either the active or passive voice, 500 only in the active voice, and 154 only in the passive voice. Among the 806 types of active verbs (306+500), 56 are used more than 15 times in the corpus data, while among the 460 types of passive verbs (306+154), 34 are used more than 15 times in the same data.

The verbs used more often in the active voice than all the other verbs are listed in Table 1. Their Cohen's *h* is larger than 0.8, except for the verb "attain."

	Active	Passive	p	Cohen's *h*
have	240	0	** p <.01	1.467
enter	127	0	** p <.01	1.377
study	84	0	** p <.01	1.347
mean	43	0	** p <.01	1.320
follow	29	0	** p <.01	1.311
play	19	0	** p <.01	1.305
learn	70	1	** p <.01	1.099
visit	47	1	** p <.01	1.032
assume	38	1	** p <.01	0.995
receive	98	5	** p <.01	0.907
reach	33	2	** p <.01	0.829
attain	25	2	** p <.01	0.755

Table 1: Verbs used more often in active voice in the data

This table includes the verbs "have" and "mean"; they are also included in the result of G&S as these verbs tend to be used in active voice. On the other hand, this table does not contain all the other verbs that tend to be used in active voice in the result of G&S, since they are not frequent enough (used only 14 times or less in either the active or passive voice) or their Cohen's *h* is not larger than 0.8.

The verbs used more often in passive voice than all the other verbs are listed in Table 2. Their Cohen's *h* is lower than -0.8.

	Active	Passive	p	Cohen's *h*
say	9	176	** p <.01	-1.609
refer	1	26	** p <.01	-1.480
believe	5	51	** p <.01	-1.294
locate	4	42	** p <.01	-1.292
bear	10	85	** p <.01	-1.277
base	2	19	** p <.01	-1.239
know	16	105	** p <.01	-1.122
bury	3	17	** p <.01	-1.068
destroy	10	39	** p <.01	-0.947
think	5	20	** p <.01	-0.939
assign	5	18	** p <.01	-0.894

Table 2: Verbs used more often in passive voice in the data

This table includes the verbs "bear" and "base"; they are also included in the result of G&S as they tend to be used in the passive voice. However, this table does not contain all other verbs that tend to be used in passive voice in the result of G&S, since they are not used frequently enough in our data (used only 14 times or less in either the active or passive voice) or their Cohen's *h* is larger than -0.8.

This table includes the verbs "believe," "think," "say," and "know" as they are used more often in

the passive voice than the active voice. This result is in contrast with the result of G&S, wherein these verbs are used more often in the active voice than the passive voice.

The verbs whose Cohen's *h* falls between -0.2 and 0.2 are listed in Table 3; they are called "neutral" verbs.

	Active	Passive	p	Cohen's *h*
call	239	182	**p < .01	-0.213
found	51	38	ns	-0.140
describe	24	17	ns	-0.108
write	55	34	ns	-0.041
name	19	11	ns	-0.007
show	21	11	ns	0.042
grant	22	11	ns	0.064
confer	24	10	ns	0.152
send	24	10	ns	0.152
put	22	9	ns	0.160
preach	18	7	ns	0.182
give	123	49	*p < .05	0.193

Table 3: Neutral verbs in the data

None of the verbs in Table 3 are included in the result of G&S.

7. Discussion

This study found that the data include verbs that are used in the active voice more often than the passive voice, and vice versa. This finding suggests that the active-passive alternation is not a purely syntactic phenomenon but rather a lexical-semantic one. The same result was obtained by G&S.

However, the list of the verbs in this study is not identical with that of G&S; although there are some similarities ("have" and "mean" in active voice and "bear" and "base" in passive voice), all the other verbs in Tables 1 and 2 are not included in their study. In addition, we can find contradictory cases between their study and ours as some verbs ("believe," "think," "say," and "know"), which are used in the active voice in their study, are used more often in the passive voice in ours.

This discrepancy is surely the result of different foci on which verbs should be considered in G&S and our study: G&S focused on all the verbs in their data, while we focused on only some frequently used verbs in our data. In addition, the corpus they used contains a variety of genres of text written by native English speakers, while our data contains definitions of terms in a limited area of interest (Buddhism) translated from Japanese into English by non-native English speakers and edited by native speakers.

It can be argued that this discrepancy between G&S and our study supports the claim that the active-passive alternation constitutes a lexico-semantic phenomenon. That is, the difference in text genres is reflected by which verbs tend to be used more often in the active voice than the passive voice. In particular, the verbs "believe," "think," "say," and "know" are used in passive voice, because their passive constructions can express situations wherein a story or incident is accepted by the general public (e.g., "it is believed that...," and "it is said that..."). Moreover, it is natural that these expressions are used more frequently than usual, as the aim of the texts in our data is to provide an introduction to a historical person or historical incident. In addition, we cannot ignore the influence of Japanese phrases that use passive voice verbs such as "*...to shinjirareteiru*" (It is believed that...), "*...to kangaerareteiru*" (It is thought that...), "*...to iwareteiru*" (It is said that...)," and "*...to shirareteiru*" (It is known that...). In future research, we aim to identify such constructions in English translations of Japanese texts that are used more often than usual (possibly) because of the influence of the original, or in English sentences produced by non-native speakers of English, such as Japanese learners of English.

The observation of these passive voice verbs with the possible influence of the original Japanese sentences seems to support the assumption mentioned in Section 4 above that English sentences translated from Japanese sentences are one of the genres of non-native English.

The observation of these passive voice verbs also seems to argue against the claim that the genre of corpus used in this study cannot be employed to address the issue of distinguishing native English and non-native English; that is, the corpus data in this study are English sentences translated from Japanese sentences with proofreading by native speakers of English, and therefore they can be less non-nativelike than other "pure" non-native English sentences, such as essays written by Japanese learners of English. However, the proofreading by native speakers of English does not necessarily render English sentences as nativelike as possible, and therefore they cannot be

"pure" native English sentences, as other types of English sentences produced by non-native speakers of English.

In this context, though, it will be productive to explore the possibility of finding more supportive results through the investigation of collexemes in the corpus data produced by non-native learners of English, with the same method as this study. This will be the goal of future research.

To support the claim that active-passive alternation constitutes a lexico-semantic phenomenon, we need to explain that some verbs can alternate between the active and passive voice without any bias toward either. G&S did not address this issue, since they only reported verbs that are distinctively biased toward the active or passive voice. As reported in Table 3, we found that some verbs are used either in the active or passive, and there is no significant difference between these two usages as far as our data is concerned. This may support the claim that active-passive alternation constitutes a syntactic phenomenon, and any bias toward the active or passive cannot be found, at least for these verbs. In this context, the behaviors of these verbs, which are found unbiased in our data, need to be investigated in different corpora or subcorpora of the same corpus, so that we may verify the possibility that these verbs can also show a tendency to be used in either the active or passive voice. This will reflect the particular characteristics of the corpus data, which will be a research question of future studies.

8. Conclusion

This study conducted a corpus-based investigation of collexemes for active-passive alternation found in the English part of an English-Japanese parallel corpus, as an attempt to use them as metrics for distinguishing native English and non-native English. The results show that some verbs in the data are used in the active voice more often than the passive voice, and vice versa, and the differences are statistically significant. However, these verbs are not the same as those found in a previous study. This fact supports the claim that active-passive alternation constitutes a lexico-semantic phenomenon that is sensitive to various factors, such as differences in genres and type of the authors of the text (e.g., native speakers vs. non-native speakers). Moreover, some verbs are neutral to the alternation, which will be addressed in future studies on the relationships between collexemes and constructions.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 17K02740.

References

Cohen, J. 2013. Statistical power analysis for the behavioral sciences (2nd ed.). Abington, UK: Routledge.

De Marneffe, M.C., & Manning, C. 2008. The Stanford typed dependencies representation. Proceedings CrossParser '08 Coling 2008: *Workshop on Cross-Framework and Cross-Domain Parser Evaluation*. Manchester, UK: Association for Computational Linguistics.

Fisher, R. A. 1922. On the interpretation of χ^2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society, *85*(1): 87–94.

Fisher, R. A. 1954. Statistical methods for research workers (12th ed.). Edinburg, UK: Oliver and Boyd.

Goldberg, A. 1995. Constructions: A construction grammar approach to argument structure. Chicago, IL: University of Chicago Press.

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

Lakoff, G. 1987. Women, fire, and dangerous things. Chicago, IL: University of Chicago Press.

National Institute of Information and Communications Technology. 2011. The Japanese-English bilingual corpus of Wikipedia's Kyoto articles, v.2.01. Retrieved from https://alaginrc.nict.go.jp/WikiCorpus/index_E.html

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

Sinclair, J. M. 1991. Corpus, Concordance, Collocation. Oxford: OUP

Tang, X. 2017. Lexeme-based collexeme analysis with DepCluster. Corpus Linguistics and Linguistic Theory, 13(1): 165–202