BA Imperative Conditional in Modern Japanese
—With a Focus on Distributional Aspects—*

SETO Yoshitaka

1. Introduction

Japanese conditionals have been extensively studied, and this study aims to examine the semantic aspect of a Japanese conditional that features a clause linkage marker, *ba*.

(1)	Sensei-ga	ire-ba	aisatsu	shinasai.
	teacher-NOM	exist-CLM	greet	do.IMP
	'If the teacher is (there), greet them.'			

The example (1) represents BA imperative conditional, where the main predicate is in the imperative mood, and the predicate in the subordinate clause is filled with *iru* in hypothetical conjugational form. The current study will refer to the position filled by the predicate in the subordinate clause as $SLOT_1$, and that in the main clause as $SLOT_2$. BA imperative conditional in Modern Japanese is recognized as having a limited distribution of lexemes in $SLOT_1$, and non-stative verbs are typically not allowed in this position (National Institute for Japanese Language and Linguistics 1964; Masuoka 1993).

The investigation of the prototypical semantic aspects of the lexemes in $SLOT_1$ and $SLOT_2$ has been limited, and the specific semantic relation between the two clauses has not been fully explored. By examining salient lexemes in each clause and characteristic combinatory patterns of lexemes, we aim to advance the descriptive work on BA imperative conditional.

In the following section, the current study attempts to identify salient lexemes in $SLOT_1$ and $SLOT_2$ to unveil specific constructional aspects of BA imperative conditional.

2. Methodology

To identify the semantic attributes of $SLOT_1$ and $SLOT_2$ in BA imperative construction and their semantic association, we employed word2vec (Mikolov, Yih, and Zweig 2013), hierarchical clustering, and collostruction analysis to identify representative lexemes for each slot.

The data for this study on BA imperative conditionals were obtained from Balanced Contemporary Corpus of Written Japanese (BCCWJ) (The National Institute for Japanese Language and Linguistics 2015) as follows: First, sentences contained in BCCWJ were tokenized using GiNZA NLP Library version 4.0. (Matsuda, Oomura, and Asahara 2019). This enabled the identification of

* This research was supported by JSPS KAKENHI Grant Number 19K13189.

the predicate in the protasis and apodosis of the conditional being studied by obtaining the dependency relation between morphemes in each sentence. After filtering instances of BA imperative conditionals, the rows in a word2vec table that corresponds to the words in $SLOT_1$ and $SLOT_2$ were extracted. Additionally, collostruction analysis was employed.

Collostruction analysis has been used extensively to identify specific representative lexemes in construction in both corpus and cognitive linguistics literature. This study employed two types of collostruction analysis: simple collexeme analysis (Stefanowitsch and Gries 2003) and covarying collexeme analysis (Stefanowitsch and Gries 2005). The former measures the strength of the collocation between an item in a construction and the construction itself, while the latter measures the strength of the collocation between two lexemes within the construction. This study employed simple collexeme analysis to identify words in $SLOT_1$ and $SLOT_2$ strongly connected with BA imperative conditionals in Modern Japanese. Furthermore, covarying collexeme analysis was conducted to determine the semantic relationship between the collexemes in $SLOT_1$ and $SLOT_2$ of the conditionals under study. The 'collostructions' R package version 0.2.0. (Flach 2021) was utilized for these tests. The lexemes in $SLOT_1$ and $SLOT_2$ were respectively defined as the collexeme for simple collexeme analysis. For example, in (2), '*tai,'* the base form of '*takere-'* in the protasis, and '*nigeru,'* the base form of '*nigero,*' were defined as the collexeme of simple collexeme analysis. In the covarying collexeme analysis, both words in $SLOT_1$ and $SLOT_2$ were defined as the collexemes simultaneously.

(2) Nige-takere-ba nigero-yo.
escape-want.IRR-CLM escape.IMP-SFP[1]
'If you want to run away, run away.' (BCCWJ LBh9_00086)

In addition, the collected collexemes in $SLOT_1$ and $SLOT_2$ were extracted from a word2vec model, chiVe (Manabe et al. 2019). A word2vec model is used to represent words as numerical vectors in a multi-dimensional space. It is typically represented as a table, where each row corresponds to a specific word, and each column represents a specific feature or aspect of the word. The fundamental premise of word2vec is that words with similar meanings are represented by a similar array of vectors in the model. This feature has made word2vec a widely-used tool in natural language processing (NLP) for a variety of applications, including text classification and machine translation. Models of word2vec are constructed from collocational information of a large text database, and chiVe was trained from NINJAL Web Japanese Corpus (NWJC). The lexemes in $SLOT_1$ and $SLOT_2$ were extracted separately from chiVe. When a lexeme in $SLOT_2$ was not contained in chiVe or too general,

[1] The abbreviations used in this paper are as follows: ADV: adverbial, COP: copula, CLM: clause linkage marker, HON: honorifics, IMP: imperative mood, NEG: negation, NMLZ: nominalizer, NOM: nominative, SFP: sentence-final particles.

a word with a similar meaning was extracted from chiVe, or a word dependent on the lexeme in $SLOT_2$ was substituted in the position.

(3) … pochi-tto shi-te-kudasai
pochi.ADV do-CLM-give.IMP
'..., please click.' (BCCWJ OY14_27547)

For example, in sentence (3), the original text contained the adverbial form of *suru* (do), *shi*, in $SLOT_2$. However, since *shi* is too general and does not accurately convey the meaning of pushing in the sentence, and *pochi-tto*, an adverb used to express the action of pushing, was selected as a better fit for $SLOT_2$. In this case, *pochi-tto* was defined as the lexeme in $SLOT_2$. The lexemes in each slot were then submitted to Rtsne (Krijthe and Van der Maaten 2022), which projected the high-dimensional data onto a two-dimensional space, enabling words with similar distributional profiles to be placed close to each other.[2] Finally, hierarchical clustering was used to visualize the relationships between the lexemes.[3]

3. Results

The retrieval of BA imperative conditional resulted in 1,195 instances with 19 types of lexemes in $SLOT_1$ and 105 types in $SLOT_2$. The result of simple collexeme analysis showed that 17 types of lexemes in $SLOT_1$ are attracted to the construction, with 13 significantly attracted lexemes. Figure 1

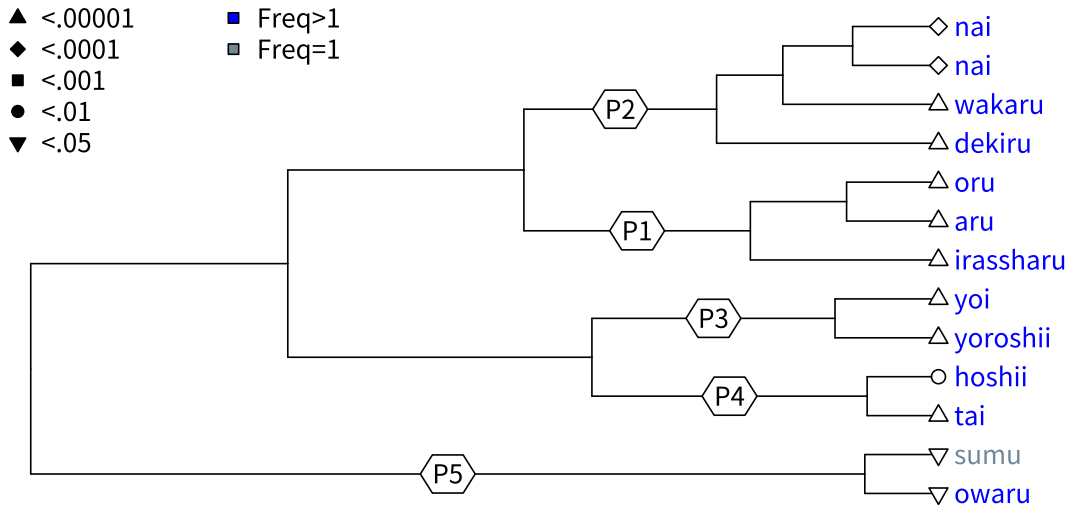

Figure 1. Lexemes in $SLOT_1$ of BA imperative conditional.

[2] The t-SNE algorithm was executed for 1,000 iterations using a perplexity parameter of 3 for the lexemes in slot1. Conversely, for slot2, 5,000 iterations were performed with a perplexity parameter of 4.
[3] The Euclidean method was used to calculate distances, and the Ward.D2 method was utilized as the clustering method for the hierarchical clustering analysis.

depicts the result of the hierarchical clustering of the significantly attracted lexemes in $SLOT_1$. The average silhouette width is highest (0.51) with five clusters, represented by P_1 through P_7. The *p*-value of the collostruction strength is displayed using font color, with a distinction made for whether the lexeme appears only once in $SLOT_1$ of BA imperative conditional.

The clusters shown in Figure 1 confirm that each cluster represents a specific semantic meaning. For example, in the case of P_5, both *sumu* and *owaru* show the meaning of the completion of an event. This cluster is distinct from the other clusters. P_4 consists of two lexemes, *hoshii* and *tai*, which typically express a person's desire. *Yoi* and *yoroshii* in P_3 both contain the meaning of a positive evaluation. The meaning of existence is observed in P_1: *Irassharu* is an honorific for a person's existence, *oru* expresses the existence of an animate being, and *aru* expresses the existence of an inanimate being. P_2 contains two lexemes, *nai* and its distinct notation, namely, 無い and ない. Although distinct in notation, their collocational distribution is similar to each other, as judged by their proximity in the dendrogram. They both express the meaning of non-existence, which is opposite in meaning to P_1. P_2 also contains two other lexemes, *wakaru* and *dekiru*, which do not share a common meaning with *nai*, and it may not be intuitive why they are in the same cluster. This is probably because a word2vec model is built from the collocational information, and *wakaru* and *dekiru* frequently co-occur with *nai*, enough for them to be placed in the same cluster. *Wakaru* and *dekiru* are semantically related to each other in that they both express an ability. Therefore, semantic features of P_2 are summarized as non-existence and ability.

Of the five identified clusters, it is suggested that the most salient cluster is P_1, as all the lexemes in $SLOT_1$ have a *p*-value lower than 0.00001. This indicates that BA imperative conditional favors the meaning of existence as the meaning of its protasis. Among the three lexemes with the meaning of existence, *aru* exhibits the highest collostructional strength. The use of *aru* in $SLOT_1$ at least has three ways of relating to its subject as demonstrated in the following examples:

(4) a. Nani-ka adobaisu-ga are-ba oshie-te-kudasai.
what-INT advice-NOM exist-CLM teach-CLM-give.IMP
'If you have some advice, please tell me.' (BCCWJ OC04_01128)

b. Moshi aru-no-de are-ba oshie-te-kudasai.
if exist-NMLZ-COP exist-CLM teach-CLM-give.IMP
'If it is the case that it exists, please tell me.' (BCCWJ OC02_03329)

c. Moshigozonji-de are-ba oshie-te-kudasai.
if know.HON-COP exist-CLM teach-CLM-give.IMP
'If it is the case that you know it, please tell me.' (BCCWJ OC02_05439)

In (4a), *aru* is used to express the hypothetical situation where certain advice exists. In this case, the subject of the lexeme corresponds to a thing. Secondly, *aru* refers to the existence of a situation in (4b). Finally, when the subject of *aru* is the hearer, it expresses the existence of a situation where the hearer is involved. The use of *aru* in (4a) accounts for 853 cases (93.22%), while the use of situation in (4b) accounts for 34 cases (3.72%), and the use of hearer in (4c) accounts for 28 cases (3.06%). These results suggest that the most typical meaning expressed in P_1 of BA imperative conditional is to express a hypothetical situation where a thing exists.

The lexemes in $SLOT_2$ with a frequency greater than one, which were attracted to BA imperative conditional, were optimally clustered into five clusters with an average silhouette width of 0.43. The results of the hierarchical clustering are presented in Figure 2, with each cluster labeled A_1 through A_5. Notably, cluster A_4, is the most significant as all the lexemes in this cluster are significantly attracted to BA imperative conditional with a *p*-value lower than 0.001. Additionally, each lexeme in this cluster expresses communication-related meanings. A_2, on the other hand, exhibits semantic

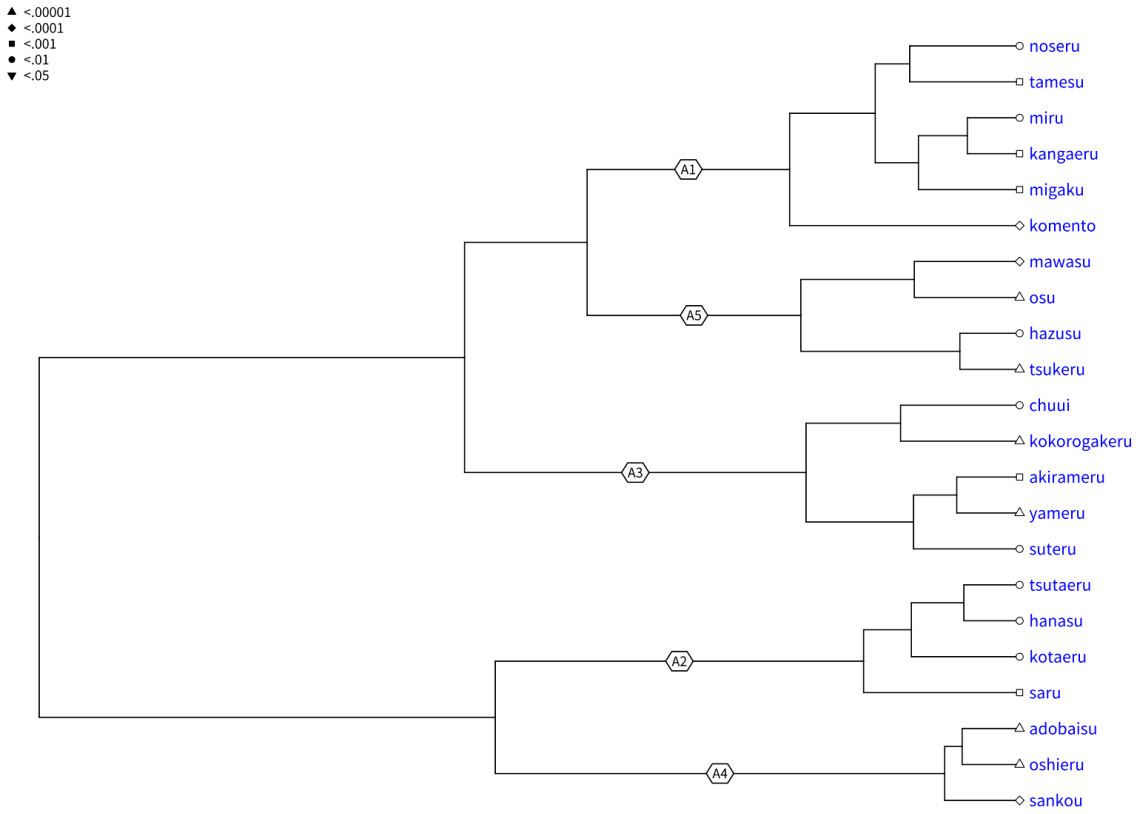

Figure 2. Lexemes in $SLOT_2$ of BA imperative conditional

similarity to A_4, as it comprises verbs of speech, except for *saru* (leave). It is suggested that the inclusion of *saru* in A_2 is due to its co-occurrence with the actions represented by the verbs in this cluster, which affects the underlying structure of the word2vec model.A_1 comprises lexemes that are related to cognitive behavior leading to specific outcomes. For example, the verb *kangaeru* expresses

the act of thinking, which can lead to the creation of an idea. Similarly, *tamesu* expresses the act of trying, which can result in the action of a trial. The verb *miru* is often used to convey the meaning of trying and seeing, as in English. *Migaku* (polishing) may not be intuitively associated with the aforementioned lexemes. However, it is frequently used to denote the meaning of improving something, such as one's skills or art of work. Consequently, the inclusion of *komento* (comment) in this cluster can be explained by considering it as a verbal output of the actions denoted by these verbs.

Cluster A_5 comprises verbs that express the manipulation of an object. Specifically, the verbs *hazusu* and *tsukeru* are associated with detaching and attaching, while *mawasu* and *osu* are related to the exercise of force onto an object.

Cluster A_3, on the other hand, relates to the direction of one's attention towards an object. The verb *chuui* (attention) and *kokorogakeru* (to note) are both convey the act of directing one's attention to an object or behavior. In contrast, the verbs *akirameru* (to give up) and *yameru* (to quit) denote the cancellation of the direction towards a particular action. The act of abandoning expressed by *suteru* is also related to the latter meaning. As we have seen, simple collostruction analysis has revealed semantically salient clusters based on word2vec model.

The covarying collostruction analysis revealed 19 types of combinations of lexemes in $SLOT_1$ and $SLOT_2$ with a frequency greater than one. The significantly attracted combination of lexemes and their respective cluster numbers are presented in Table 1where the columns $SLOT_1$ and $SLOT_2$ cluster represent the clusters identified in the simple collostruction analysis.

$SLOT_1$	$SLOT_2$	$SLOT_1$ cluster	$SLOT_2$ cluster	collostruction strength
aru	*oshieru*	P_1	A_4	81.46
nai	*saru*	P_2	A_2	32.15
nai	*yameru*	P_2	A_3	24.58
yoroshii	*komento*	P_3	A_1	12.71
yoi	*sankou*	P_3	A_4	12.1
nai	*hazusu*	P_2	A_5	12.1
nai	*suteru*	P_2	A_3	12.1
nai	*migaku*	P_2	A_1	12.1
nai	*akirameru*	P_2	A_3	12.1
nai	*tsutaeru*	P_2	A_2	11.74
nai	*kokorogakeru*	P_2	A_3	10.41
nai	*hanasu*	P_2	A_2	10.05
nai	*osu*	P_2	A_5	8.91
dekiru	*mawasu*	P_2	A_5	8.6
nai	*tsukeru*	P_2	A_5	8.39
nai	*kangaeru*	P_2	A_1	6.6
nai	*kangaeru*	P_2	A_1	6.26
nai	*miru*	P_2	A_1	6.26
irassharu	*oshieru*	P_1	A_4	4.12

Table 1. Covarying collexemes of $SLOT_1$ and $SLOT_2$ of BA imperative conditional.

The result of the covarying collexeme shows that the significantly attracted patterns of BA imperative conditional are categorized into specific semantic relations. Among these patterns, the most salient ones include the lexemes *aru* in $SLOT_1$ and *oshieru* in $SLOT_2$. An instance of these lexemes in their respective slots can be seen in the following sentence (5):

(5) Yoi renshuu houhou -ga are-ba oshie-te-kudasai.
good practice method-NOM exist-CLM teach-TE-give.IMP
'If there is a good method for practice, please teach me.' (BCCWJ OC01_01864)

The observed pattern highlights the salient semantic relation between the existence of a situation in the protasis and the request for offering information in BA imperative conditional. This semantic relation is further supported by the occurrence of *irassharu* in the same cluster as *aru* (P_1) and *oshieru* as a significant pattern.

The aforementioned semantic relation is also observed with other lexemes that have similar semantic content. As previously demonstrated, the lexemes in A_1 are semantically linked to those in A_2 in terms of the transmission of information. This observation can be extended to P_1 and P_2, which have opposite meanings of existence and non-existence, respectively. The examples provided in (6ab) exemplify this pattern with the request for transmitting information in its apodosis using lexemes in A_2.

(6) a. Kare-ga kii-te-i-nakere-ba kou tsutae-te-kudasai.
3SG-NOM listen-CLM-exist-NEG-CLM this tell-TE-give.IMP
'If he is not listening, please say this to him.' (BCCWJ LBo9_00200)

b. Moshi sashitsukae nakere-ba hanashi-te-kudasai.
if obstacle exist.NEG-CLM tell-TE-give.IMP
'If it doesn't bother you, please tell me.' (BCCWJ OB1X_00087)

In sentences (6ab), the covarying pattern with the most frequent lexeme in $SLOT_1$ is *nai*, which appears in 14 out of 19 covarying combination types in P_2. This lexeme co-occurs with communication verbs such as *tsutaeru* and *hanasu* in A_2, similar to those in A_4 that contain *oshieru*. This semantic similarity between the protasis and apodosis suggests that the combination of lexemes in P_1 and P_4, and those in P_2 and A_2, are in close semantic proximity. These instances confirm that the pattern of hypothesizing the state of affairs where a particular entity does not exist in the protasis and the order of offering information in the apodosis is favored in BA imperative conditional.

Another salient pattern in BA imperative conditional is when a speaker orders their interlocutor to direct or divert their attention to or from a certain action by suggesting them to think or forget

about it. This pattern is expressed through the combination of lexemes SLOT$_1$ and A$_1$ or A$_3$, which may include *kangaeru*, *miru*, *kokorogakeru* and *akirameru*. In such cases, the non-occurrence of an event is expressed using the negative particle *nai*. Another semantic relation observed with *nai* in SLOT$_1$ of BA imperative conditional is when it pertains to performing a manipulation, which is represented by the lexemes in A$_5$.

The last salient pattern in BA imperative conditional is hypothesizing the hearer's ability or the convenience of the speaker to perform the action requested in the apodosis. When the speaker hypothesizes the hearer's ability in the protasis, the lexeme *dekiru* is typically used. On the other hand, the lexeme *yoi* (good) occupies SLOT$_1$ to hypothesize the convenience of the hearer. In such cases, the prominent lexemes in A$_4$ and A$_1$ are *sankou* and *komento,* respectively. Although they belong to different clusters, they share a common trait in that they are related to the interaction of information.

4. Discussion and Conclusion

The current study has investigated the semantic properties of BA imperative conditional focusing on the lexemes in SLOT$_1$ and SLOT$_2$ of BA imperative conditional and their combinatory patterns. The result has shown that particular lexemes are significantly attracted to the conditional under study. It has been shown that some lexemes are salient in the protasis, and others are in the apodosis. In other cases, particular combinations of the lexemes are salient in each slot.

The study identified salient patterns of BA imperative conditionals that suggest a multi-level construction network. According to Hilpert (2019: 21), one way to determine if a pattern is an instance of a construction is to look for frequent collocational patterns. This study found that the attracted lexemes and their combinatory patterns identified with two types of collostruction analysis, indicating that it can be considered a construction. The findings suggest that the existence of specific sub-constructions of BA imperative conditionals, which vary in specificity and are connected with semantic similarities between each construction. The specificity of a construction is lower when a single slot is filled with a lexeme than when both slots are occupied with specific lexemes. Therefore, this information can be obtained from the results of simple and covarying collostruction analysis, respectively. Figure 3 partially diagrams the constructions identified in this study.

The construction network diagrammed in Figure 3 illustrates the vertical relation between each node to show the degree of specificity of the construction. The uppermost node represents the most general level of BA imperative conditional, while the lower nodes represent increasingly specific subconstructions. The protasis is colored in red and expresses the meaning of *condition(cond.)*, while the apodosis is boxed in blue and expresses the meaning of *order or request (o/r)*.

At the lowest node, we find the most specific subconstruction of BA imperative construction, which is identified by two slots in the protasis and apodosis filled with two lexemes, *are* and *oshiete*, as identified by covarying collexeme analysis. Above this level, the constructions identified with

simple collostruction analysis have SLOT$_1$ or SLOT$_2$ filled with a specific lexeme. For example, the nodes with *are* and *irasshare* in SLOT$_1$ represent a subconstruction of BA imperative conditional.

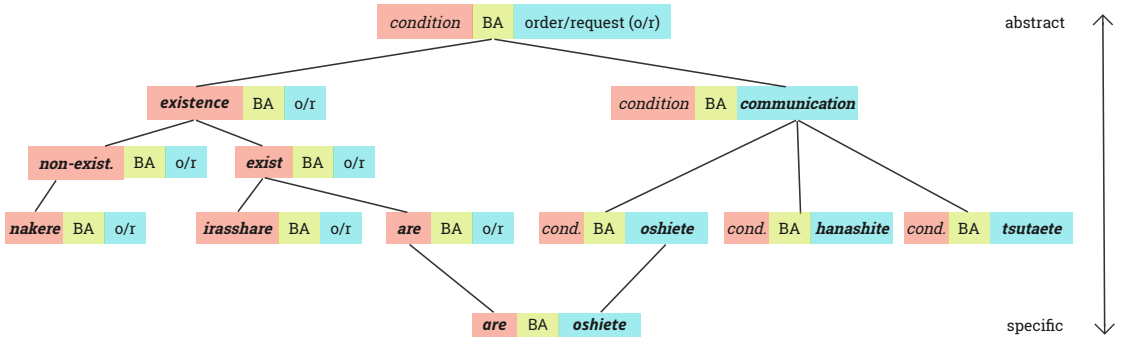

Figure 3. Partial construction network of BA imperative conditional.

The connections between nodes indicate the semantic schematization of the lower constructions. For instance, the pattern of '*exist* BA order/request' schematizes the lexemes in SLOT$_1$ of '*irasshare/are* BA order/request.' Also, the construction at a more abstract level, '*existence* BA order/request' abstracts '*non-exist* BA order/request' and '*exist* BA order/request.' These correspond to the semantic clusters identified through the hierarchical clustering of the lexemes in SLOT$_1$ and SLOT$_2$ based on the word2vec model.

The identified salient clusters of lexemes in SLOT$_1$ in the protasis are categorized into four semantic groups: existence (P$_1$ and part of P$_2$), desirability (P$_3$ and P$_4$), possibility (part of P$_2$), and completion (P$_5$). The first three semantic group are related to each other in that they all represent the state, while the fourth group refers to the non-state. As Figure 1 illustrates, the lexemes in this group are displaced from the other groups, and it confirms this semantic distinctness. Meanwhile, the lexemes in SLOT$_2$ are classified into five types, namely, analysis (A$_1$), manipulation (A$_5$), attention (A$_3$), and communication (A$_2$ and A$_4$). The last cluster is also distinct from the other clusters, and it shows the conceptual distinctness of the cluster. Figure 2 depicts that the last semantic group is distant from the other groups. This can be accounted for by the degree of intersubjectivity. In the last semantic group, the events that the verbs typically represent require at least two people in the usage event, the speaker and the hearer. Conversely, the other groups do not require interlocutors. Hierarchical clustering successfully captures the conceptual distinctness in the lexemes in SLOT$_1$ and SLOT$_2$.

The results of this study demonstrate the effectiveness of utilizing word2vec modeling, hierarchical clustering, and collostruction analysis to identify significant semantic aspects of BA imperative conditional, with a particular emphasis on the lexemes in SLOT$_1$ and SLOT$_2$. Despite the success of this approach, certain issues must be considered. Notably, the hierarchical clustering method based on tSNE outcomes may not consistently produce semantically intuitive clusters. For example, P$_1$

comprises lexemes conveying the idea of existence, yet *nai*, which is semantically linked to this sense, is assigned to P_2. In such instances, it is essential to scrutinize the interpretation of the clustering. Despite these potential difficulties, the overall findings of this study are promising and can be extended to future investigations exploring other conditional types.

References

Flach, Susanne (2021) Collostructions: An R Implementation for the Family of Collostructional Methods. R. www.sfla.ch.

Hilpert, Martin (2019) *Construction Grammar And Its Application To English*. 2nd ed, Cheshire: Edinburgh University Press.

Krijthe, Jesse, and Laurens Van der Maaten (2022) Rtsne: T-Distributed Stochastic Neighbor Embedding Using a Barnes-Hut Implementation. R.

Manabe Youshun, Oka Teruaki, Umikawa Hiroki, Takaoka Kazuma, and Asahara Masayuki (2019) Fukusuu ryuudo-no bunkatsu kekka-ni motozuku nihongo tango bunsan hyougen. *Proceedings of the Twenty-fifth Annual Meeting of the Association for Natural Language Processing*, 1407–10.

Masuoka Takashi (1993) Nihongo-no joukenhyougen-ni tsuite. In Masuoka Takashi (eds.) *Nihongo-no joukenhyougen*, 1–20, Tokyo: Kurosio Publishers.

Matsuda, Hiroshi, Mai Oomura, and Masayuki Asahara (2019) Tan Tan'i Hinshi-No Youhou Aimaisei Kaiketsu-to Izon Kankei Raberingu-No Douji Gakushuu. *Proceedings of the Twenty-Fifth Annual Meeting of the Association for Natural Language Processing*, 201–4.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013) Linguistic Regularities in Continuous Space Word Representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–51, Atlanta, Georgia: Association for Computational Linguistics. https://aclanthology.org/N13-1090.

National Institute for Japanese Language and Linguistics (1964) *Gendai Zasshi Kyuujusshu No Yougo Youji: Dai 3 Bunsatsu (Bunseki)*. NINJAL report, Shuuei Shuppan.

Stefanowitsch, Anatol, and Stefan Th. Gries (2003) Collostructions: Investigating the Interaction of Words and Constructions. *International Journal of Corpus Linguistics* 8**(2)**: 209–43. https://doi.org/10.1075/ijcl.8.2.03ste.

——— (2005) Covarying Collexemes. *Corpus Linguistics and Linguistic Theory* 1**(1)**: 1–43. https://doi.org/10.1515/cllt.2005.1.1.1.

The National Institute for Japanese Language and Linguistics (2015) Balanced Corpus of Contemporary Written Japanese., Tokyo.