Extracting Topic-Specific Collocations from J-TOCC2 and Their Application to Japanese Language Education

Tzuhsuan MA, Naoki NAKAMATA
Kyoto University of Foreign Studies, The University of Osaka
kennji.ma@gmail.com, n.nakamata.ciee@osaka-u.ac.jp

Abstract

This study extracts topic-specific collocations from the Japanese Topic-Oriented Conversation Corpus, Phase 2 (J-TOCC2), and evaluates their value for Japanese language education. Using lemma information from the corpus's long-unit-word analysis, we extracted trigrams matching the pattern noun + case-marking particle + verb and compiled frequency tables for each of the 16 Sustainable Development Goal topics. For each topic, log-likelihood ratios (LLRs) were calculated against the combined data from the other 15 topics. Sequences with an LLR of at least 10.83 ($p < .001$) and a corpus-wide frequency of at least five were treated as topic-specific candidate collocations. The analysis identified semantically relevant expressions such as *fūdorosu o herasu* 'reduce food loss' for Zero Hunger and *infura ga totonou* 'infrastructure is developed' for Clean Water and Sanitation. A total of 296 candidate collocations were extracted across the 16 topics, with the number of candidates ranging from 10 to 39 per topic. A pedagogical evaluation of 89 candidates from the first five topics classified 53.9% as high-priority items, 34.8% as basic combinations, and 11.2% as context-dependent fragments. These findings show that J-TOCC2 can support the empirical selection of topic-specific collocations for syllabus and materials development.

Keywords: topic-specific collocations, J-TOCC2, log-likelihood ratio, Japanese language education, corpus-based pedagogy

1. Introduction

Language education, including Japanese language education, has increasingly shifted from knowledge-accumulation models based on grammatical patterns toward task-based and other action-oriented approaches. This shift reverses the conventional direction of syllabus design: rather than teaching a predetermined inventory of forms and subsequently asking learners to use them, instructors must first identify the vocabulary and expressions required for a particular task and then provide targeted instruction. Discussing a specified topic can itself be treated as a communicative task, especially in intermediate and advanced courses where learners are expected to engage with social issues.

Previous corpus-based studies of topic variation have often focused on individual keywords. However, successful participation in topic-based discussion also depends on recurrent multiword combinations, such as those following the pattern noun + case-marking particle + verb. Because such combinations are larger than individual words and often conventional rather than fully predictable, learners may understand every component word but still fail to produce the combination naturally. Systematic identification of topic-specific collocations is therefore needed for materials development and classroom instruction.

This study uses the Japanese Topic-Oriented Conversation Corpus, Phase 2 (J-TOCC2), to extract statistically distinctive collocation candidates for each of its 16 Sustainable Development Goal (SDG) topics. It then evaluates how useful the extracted candidates are for Japanese language education. The study addresses two questions: (1) what combinations following the pattern noun + case-marking particle + verb are statistically characteristic of individual SDG topics, and (2) what proportion of those combinations merit explicit pedagogical attention?

2. Related Work

2.1 Extraction of Topic-Specific Words

Representative studies of domain- or topic-specific vocabulary include Uchiyama et al. (2004), Matsushita (2011), and Nakamata and Ma (2022). Uchiyama et al. compared several indices, including the log-likelihood ratio and the Dice coefficient, for selecting domain-specific words for English language education. Matsushita extracted Japanese academic common vocabulary, whereas Nakamata and Ma examined topic-specific vocabulary in J-TOCC. Although their target vocabularies differ, both studies used the log-likelihood ratio as a measure of keyness. These studies show that statistical comparison between a target corpus and an appropriate reference corpus is an effective way to identify characteristic lexical items.

2.2 Extraction of Collocations

Studies of collocation extraction include Tanomura (2009) and Lee and Sasaki (2015). Tanomura compared three methods using a purpose-built web corpus: co-occurrence analysis at the word level, co-occurring word-sequence analysis using word-level n-grams, and character-level n-gram analysis. The study concluded that word-level n-grams were the most suitable method for extracting collocational information. Lee and Sasaki used n-gram frequencies from a textbook corpus to construct pedagogically graded collocation data for Japanese language education.

In summary, previous research has used keyness statistics such as the log-likelihood ratio to extract characteristic words and has used n-grams to identify collocations. However, comparatively few studies have combined the two approaches by applying a keyness statistic directly to lemma-based n-gram frequencies. The present study therefore calculates topic-wise n-gram frequencies and applies the log-likelihood ratio to extract topic-specific collocation candidates.

3. Data and Methods

3.1 J-TOCC2

The linguistic resource used in this study is J-TOCC2, the second phase of the Japanese Topic-Oriented Conversation Corpus. J-TOCC2 contains five-minute conversations in which pairs of native Japanese university students who already knew each other discussed each of 16 SDG topics. The same pair discussed all topics for the same amount of time, and participant composition was controlled for region and gender. Each topic contains 120 conversation files, or approximately ten hours of speech, making the 16 subcorpora directly comparable while minimizing variation unrelated to topic.

J-TOCC2 is suitable for the present analysis for two reasons. First, its 16 parallel topic subcorpora can alternately serve as target and reference corpora, which facilitates the extraction of statistically distinctive expressions. Second, the data consist of spontaneous speech elicited from Japanese-speaking undergraduates rather than expressions invented by researchers or drawn from translated teaching materials. The corpus can therefore provide authentic models of how university students discuss social issues.

3.2 Extraction Procedure

We used lemma information in the long-unit-word morphological analysis distributed with J-TOCC2. From each topic subcorpus, we extracted lemma trigrams matching the part-of-speech pattern noun + case-marking particle + verb and compiled a frequency table. The analysis focused on this pattern because it captures many pedagogically important predicate-argument combinations while remaining sufficiently frequent for topic-wise statistical comparison.

For each of the 16 topics, the target subcorpus was compared with a reference corpus formed by combining the other 15 topics. We calculated the log-likelihood ratio (LLR; Dunning, 1993) for every trigram. A sequence was classified as a topic-specific candidate collocation when its LLR was at least 10.83, corresponding to $p < .001$, and its total frequency across all 16 topics was at least five.

3.3 Pedagogical Evaluation

To evaluate pedagogical usefulness, we examined all 89 candidates extracted from the first five topics: No Poverty, Zero Hunger, Good Health and Well-Being, Quality Education, and Gender Equality. The central criterion was whether a learner who knew the component words would nevertheless have difficulty producing the combination independently. Candidates were assigned to three categories: high-priority, basic, and context-dependent.

High-priority items were conventional or semantically abstract combinations whose form is difficult to infer directly from the meanings of their components. Basic items were transparent combinations involving concrete nouns or everyday actions and were therefore assigned a lower instructional priority. Context-dependent items were incomplete n-gram fragments that could not be taught appropriately without reconstructing a larger phrase or clause from the concordance context.

4. Results and Discussion

4.1 Number of Extracted Collocations

Figure 1 presents the number of candidate collocations extracted at the $p < .001$ level for each topic.

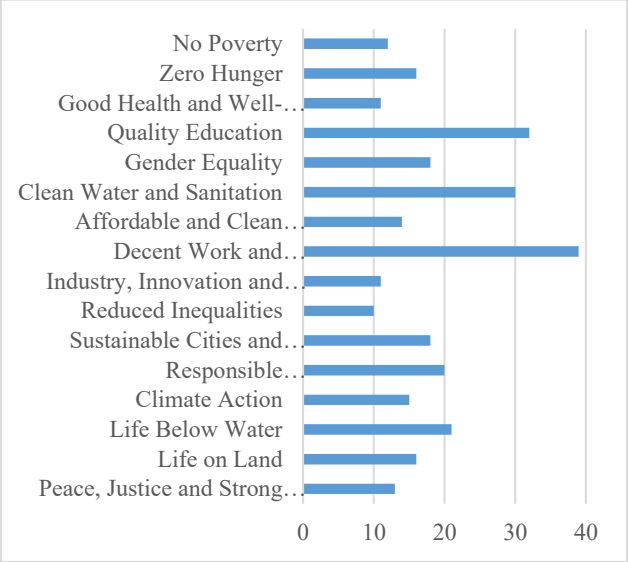

Figure 1: Number of topic-specific candidate collocations

Table 1 presents representative topic-specific collocations extracted for each of the 16 SDG topics in J-TOCC2.

Topic	Japanese collocation	Translation
No Poverty	*kakusa ga umareru*	disparities emerge
Zero Hunger	*fūdorosu o herasu*	reduce food loss
Good Health and Well-Being	*chiryō o ukeru*	receive treatment
Quality Education	*mochibe o ageru*	increase motivation
Gender Equality	*me o mukeru*	turn one's attention to
Clean Water and Sanitation	*infura ga totonou*	infrastructure is developed
Affordable and Clean Energy	*mina ni ikiwataru*	reach everyone
Decent Work and Economic Growth	*hatarakigai ga aru*	find one’s work rewarding
Industry, Innovation and Infrastructure	*kiban o tsukuru*	build a foundation
Reduced Inequalities	*kachi o oku*	place value on
Sustainable Cities and Communities	*machizukuri o suru*	engage in community development
Responsible Consumption and Production	*kankyō ni hairyosuru*	take the environment into consideration
Climate Action	*te o kuwaeru*	make modifications
Life Below Water	*umi ni mensuru*	face the sea
Life on Land	*ki ga haeru*	trees grow
Peace, Justice and Strong Institutions	*hanashiai de kaiketsusuru*	resolve through discussion

Table 1: Representative Topic-Specific Collocations

A total of 296 candidate collocations were extracted across the 16 topics. The number of candidates differed substantially by topic, ranging from 10 to 39. Decent Work and Economic Growth yielded the largest set (39), whereas Reduced Inequalities yielded the smallest (10). This difference appears to reflect both the lexical breadth of a topic and the degree to which its expressions overlap with other SDG topics. Decent Work and Economic Growth invites discussion of employment, wages, working conditions, careers, and macroeconomic activity, producing several relatively distinct predicate-argument combinations. Reduced Inequalities is broader and shares much of its vocabulary with poverty, gender, education, and work; consequently, fewer combinations are contrastively distinctive when the other 15 topics are used as the reference corpus. This interpretation remains tentative because the number of candidates is also affected by frequency distributions and the selected threshold.

The extraction nevertheless produced combinations closely related to the content of individual topics. Examples include *fūdorosu o herasu* 'reduce food loss' for Zero Hunger and *infura ga totonou* 'infrastructure is developed' for Clean Water and Sanitation. Such expressions represent not merely frequent words but reusable form-meaning pairings that can support topic-based speaking activities.

4.2 Pedagogical Evaluation of the Extracted Collocations

Table 2 summarizes the three-level evaluation of the 89 candidates from the first five topics. High-priority collocations accounted for 53.9% of the candidates. Basic combinations with a relatively low instructional priority accounted for 34.8%, and context-dependent fragments accounted for 11.2%.

Topic	High-priority	Basic	Context-dependent	Total
No Poverty	9 (75%)	2 (16.7%)	1 (8.3%)	12
Zero Hunger	10 (62.5%)	5 (31.3%)	1 (6.3%)	16
Good Health and Well-Being	5 (45.5%)	5 (45.5%)	1 (9.1%)	11
Quality Education	16 (50%)	14 (43.8%)	2 (6.3%)	32
Gender Equality	8 (44.4%)	5 (27.8%)	5 (27.8%)	18
Overall	48 (53.9%)	31 (34.8%)	10 (11.2%)	89

Table 2: Pedagogical evaluation of collocations from the first five topics

Tables 3-5 provide representative examples from each category.

Topic	Japanese collocation	Translation
No Poverty	*seikatsu hogo o ukeru* *kakusa ga umareru*	receive public assistance disparities emerge
Zero Hunger	*fūdorosu o herasu* *yunyū ni tayoru*	reduce food loss rely on imports
Good Health and Well-Being	*kōreika ga susumu* *chiryō o ukeru*	population aging progresses receive treatment
Quality Education	*mochibe o ageru* *kyōiku ga ikitodoku*	increase motivation education reaches everyone
Gender Equality	*ue ni tatsu* *me o mukeru*	hold a position of authority turn one's attention to

Table 3: Examples classified as high-priority

Topic	Japanese collocation	Translation
No Poverty	*okane ga aru* *kodomo ga iru*	have money have children
Zero Hunger	*gohan o taberu* *shashin o toru*	eat a meal take a photograph
Good Health and Well-Being	*byōin ni iku* *isha ga iru*	go to a hospital there are doctors
Quality Education	*gakkō ni iku* *kyōiku o suru*	go to school provide education
Gender Equality	*shufu ni naru* *seijika ni naru*	become a homemaker become a politician

Table 4: Examples classified as basic

Topic	Japanese collocation	Translation
No Poverty	*hinkon to iu*	'poverty,' as it is called ...
Zero Hunger	*kiga to iu*	'hunger,' as it is called ...
Good Health and Well-Being	*iryō ga ukeru*	[fragment involving receiving medical care]
Quality Education	*mono ga chigau*	the things are different ...
Gender Equality	*tokoro de iu*	speaking in terms of ...

Table 5: Examples classified as context-dependent

High-priority items such as *kakusa ga umareru* 'disparities emerge,' *kōreika ga susumu* 'population aging progresses,' and *kyōiku ga ikitodoku* 'education reaches everyone' combine abstract nouns with verbs that show conventional selectional preferences. Learners may understand the intended concepts yet produce semantically plausible but less conventional alternatives such as *kakusa ga arawareru* 'disparities appear,' *kōreika ni naru* 'become population aging,' or *kyōiku ga hirogaru* 'education spreads.' Explicit instruction is therefore valuable because it directs learners toward combinations preferred in actual discussion.

By contrast, basic items such as *kodomo ga iru* 'have children,' *gohan o taberu* 'eat a meal,' and *byōin ni iku* 'go to a hospital' typically contain concrete nouns and everyday actions. Their components are loosely combined and their meanings are transparent, so learners who know the individual words can often construct them through ordinary semantic association. They remain useful as topic vocabulary, but they have a lower priority for focused collocation instruction.

The context-dependent category includes sequences such as *hinkon to iu*, *mono ga chigau*, and *tokoro de iu*. These are fragments cut from larger constructions, for example *hinkon to iu mondai* 'the problem called hunger' or a preceding modifier followed by *mono ga chigau*. Because a fixed trigram window does not necessarily coincide with a complete syntactic or semantic unit, such items cannot be adopted directly as teaching targets without checking their

concordance lines and reconstructing the surrounding context.

Overall, the results indicate that statistical extraction is effective as a candidate-generation procedure but is not sufficient for direct pedagogical adoption. Approximately half of the candidates were strong targets for explicit instruction, whereas the remainder required either lower prioritization or contextual reconstruction. Combining keyness statistics with qualitative pedagogical screening therefore offers a practical workflow for building topic-specific teaching lists.

5. Conclusion and Future Work

This study applied the log-likelihood ratio to lemma-based n-grams in J-TOCC2 and extracted sequences following the pattern noun + case-marking particle + verb that were characteristic of individual SDG topics. The procedure identified pedagogically relevant collocations, including *fūdorosu o herasu* 'reduce food loss' and *infura ga totonou* 'infrastructure is developed.' The pedagogical evaluation further showed that 53.9% of the candidates from the first five topics were high-priority combinations that learners would be unlikely to produce solely from knowledge of the component words.

The method also extracted transparent everyday combinations and incomplete trigram fragments. These findings do not negate the success of extraction; rather, they show that statistical distinctiveness and pedagogical priority are related but non-identical dimensions. Candidate lists should therefore be accompanied by concordance checking, contextualized examples, and a pedagogical rating.

Several issues remain for future research. First, the manual evaluation should be extended to all 16 topics and conducted by multiple raters so that inter-rater reliability can be established. Second, the usefulness of the selected collocations should be validated with learner production data and classroom materials. Third, other patterns such as noun + no + noun and adjective + noun should be examined. Because n-gram frequencies are lower than single-word frequencies, some part-of-speech patterns may not provide enough data for stable topic-wise statistical analysis; alternative thresholds and association measures will therefore need to be compared. Addressing these issues will help transform corpus-derived candidate lists into reliable resources for topic-based Japanese language education.

6. Acknowledgements

This study was supported by JSPS KAKENHI Grant Number 23K21940.

7. Bibliographical References

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

Lee, J. and Sasaki, K. (2015). Kyookasho koopasu o riyoo shita nan'ido-betsu korokeeshon jisho no teean [A proposal for a proficiency-graded collocation dictionary using a textbook corpus]. *Dai 8-kai Koopasu Nihongogaku Waakushoppu Yokooshuu,* pp. 273-278.

Matsushita, T. (2011). Nihongo no gakujutsu kyootsuu goi (akademikku waado) no chuushutsu to datoosei no kenshoo [Extracting and validating Japanese academic common vocabulary]. *2011-nendo Nihongo Kyooiku Gakkai Shunki Taikai Yokooshuu,* pp. 244-249.

Nakamata, N. and Ma, T. (2022). 'Nihongo Wadai-betsu Kaiwa Koopasu: J-TOCC Goi-hyoo' no kookai to Nihongo kyooiku-muke joohoo saito ni muketa shihyoo no kentoo [Release of the J-TOCC vocabulary list and examination of indices for a Japanese-language education information website]. *Gengo Shigen Waakushoppu 2022 Happyoo Ronbunshuu,* pp. 53-58.

Tanomura, T. (2009). Koopasu kara no korokeeshon joohoo chuushutsu: Bunseki shuhoo no kentoo to korokeeshon jiten koomoku no shisaku [Extracting collocational information from corpora: An examination of analytical methods and a trial compilation of collocation dictionary entries]. *Handai Nihongo Kenkyuu, 21,* pp. 21-41.

Uchiyama, M., Chujo, K., Yamamoto, E., and Isahara, H. (2004). Eigo kyooiku no tame no bun'ya tokuchoo tango no sentei shakudo no hikaku [A comparison of measures for selecting domain-specific words for English language education]. *Shizen Gengo Shori, 11(3),* pp. 165-197.

8. Web Resource

Nakamata, N. (2026). Japanese Topic-Oriented Conversation Corpus, Phase 2: J-TOCC2. http://nakamata.info/database/ (accessed August 4, 2026).