Semantic and textual characteristics of the colligational framework *the N1 of the N2* in argumentative essays by L1 speakers of English and Japanese

Joe Geluso

Abstract・はじめに

Discontinuous formulaic language refers to recurrent sequences of words with one or more variable slots. Examples include *on the * hand* and *the * of the ** where the asterisks represent variable slots. These sequences of words have also been called “frameworks.” Previous research has suggested that frameworks, despite their variability, may attract internal constituents that are semantically similar. The present study uses corpus and computational approaches to achieve the goal of better understanding the semantic characteristics of the recurrent framework *the N1 of the N2* where N1 and N2 represent nouns occupying the variable slots in an otherwise fixed sequence. A corpus of English argumentative essays representing first language (L1) speakers of English and Japanese was used to compare how the two groups make use of the framework. Specifically, the group of N1s used by each L1 group was compared using network analysis to better understand the semantic groupings and relationship between the N1s. A second goal was to analyze the framework’s role in contributing to the texture (i.e., cohesion and coherence) of a text. Findings suggest that the set of N1s used by the L1 English authors are more semantically interrelated

and qualitatively different from the set of N1s used by the L1 Japanese authors. The L1 English authors more frequently use abstract N1s compared to the L1 Japanese authors who more frequently use concrete N1s. The two groups also differ in how they use the framework to contribute to the texture of a text: the L1 Japanese authors use more direct repetition of the N2 throughout their essays than the L1 English authors. Pedagogical implications include more explicit teaching of the framework, including the nature of the nouns that fill the variable slots, and the textual functions that the framework typically serves.

1. Introduction

Phraseology, or the patterning of language, permeates language at all levels: from the pairing of individual words to patterns across an entire text. Firth's (1957) widely-cited statement "You shall know a word by the company it keeps" (p. 11) is motivated by the role collocation, or words that co-occur more frequently than chance would predict, plays in the meaning of a particular word. Firth famously exemplified collocation with the example of *powerful* and *strong*, where the former co-occurs more frequently with *car* and the latter with *tea*. Since Firth's seminal work, much research has been done on the patterning of language. Sinclair (1991), for instance, proposed the "principle of idiom" or that semi-preconstructed phrases are as readily available in the mind of a speaker as individual words, and comprise much of the language that speakers regularly use.

Much work on formulaic language has been done under the umbrella of "lexical bundles." The term "lexical bundle" first appeared in Biber et al. (1999) and was defined as "recurrent expressions, regardless of the idiomaticity, and regardless of their structural status" (p. 990). Examples of

frequent 4-word lexical bundles in academic registers are *as a result of* and *the nature of the*. Studies on lexical bundles have often aimed to elucidate the types of discourse functions bundles fulfill within different text types. For example, numerous studies have examined how bundles express stance (e.g., *it is important to*), are used as discourse organizers (e.g., *on the other hand*), or fulfill referential functions (e.g., *the nature of the*) (Biber et al., 2004). While lexical bundles focus on *continuous* sequences of words, other types of formulaic language represent *discontinuous* sequences of words, or words with a variable slot (Eeg-Oloffson & Altenberg, 1994; Renouf & Sinclair, 1991).

To describe discontinuous formulaic language, Renouf and Sinclair (1991) used the term "collocational framework" and defined it as a "discontinuous sequence of two words, positioned at one word remove[sic] from each other" (p. 128). Renouf and Sinclair, using spoken and written sub-corpora from the Birmingham Collection of English Text, focused on frameworks made up of grammatical words, such as *a * of* and *for * of*, where the asterisk represents a variable slot. They found that frameworks were more frequent in written than spoken texts, and claimed that "[c]o-occurrences in the language most commonly occur among grammatical words" (p. 128). The framework *the * of the* in particular has consistently been found to be the most frequent frame in both written and spoken registers (Biber, 2009; Garner, 2016; Gray & Biber, 2015; Hasselgård, 2019; Römer, 2010). This framework is of interest given its high frequency and inclusion of a noun phrase and an *of*-phrase: two features that have been found to be extremely common in academic writing given their high information density properties (Biber & Gray, 2016; Hyland, 2008).

1.1. Frameworks and semantics

Not only are frameworks comprised of grammatical words frequent, researchers have observed that such frameworks exhibit a tendency to enclose words that feature similar semantic characteristics (Marco, 2000; Renouf & Sinclair, 1991; Eeg-Olofsson & Altenberg, 1994). For instance, Eeg-Olofsson and Altenberg (1994) explained that abstract nouns such as *trouble, problem, question,* and *situation* were likely to fill the frame *the * is*. Marco (2000) argued that the framework *a * of* often attracts collocates that express quantity or measure such as *number, percentage,* etc. and nominalizations that can be quantified such as *an accuracy of 55%* or *a specificity of 90%*.

The idea of semantically similar fillers in frameworks is also investigated in the present paper with respect to the recurrent framework *the * of the **. Hasselgård (2016 & 2019) investigated this framework using the nomenclature *the N1 of the N2* where N1 and N2 stand for "Noun 1" and "Noun 2," respectively. For example, *the end of the war* and *the rejection of the possibility* are realizations of the framework *the N1 of the N2*. Hasselgård (2019) called this sequence a "colligational framework" as an adaption of Renouf and Sinclair's (1991) "collocational framework." Hasselgård (2019) observed that while collocation involves the co-occurrence of words, colligation involves the co-occurrence of words and specific grammatical patterns. In the framework *the N1 of the N2*, the lexical items are attracted to the grammatical categories of the noun and *of*-phrases that are inherent to the framework. Hasselgård (2019) specifically focused on the semantic relationship between the N1 and N2 in English academic writing by L1 (first language) speakers of English compared to L1 speakers of Norwegian. The present study also investigates the semantic profiles of the fillers, but focusses more heavily on the

semantic similarity of the N1s in relation to each other in argumentative essays by L1 speakers of English and Japanese. In addition to analyzing the semantic similarity of fillers within the framework, the framework in its entirety is analyzed for its overall role in creating texture, or cohesion and coherence throughout a text.

1.2. Frameworks and texture

Formulaic language plays a role in the construction of discourse, and hence has a role in the creation of texture in a text (Biber et al., 2004; Gray & Biber, 2015; Nesi & Basturkmen, 2006). To further explore this idea, the term "texture" must be defined and operationalized.

Halliday and Hasan (1976) proposed that an individual text is best regarded as a semantic unit and that "a text has texture… it derives this texture from the fact that it functions as a unity with respect to its environment" (p. 2). According to Halliday and Hasan, texture is created by cohesive relations within a text, with different resources functioning to create texture (p. 2). Hoey (1991) argued that "lexical cohesion is the dominant mode of creating texture… and the study of cohesion in text is to a considerable degree the study of patterns of lexis in text" (p. 10). The following paragraphs will describe three textual resources to which patterned language contributes in order to create texture in a text: information structure, connectives, and reiteration.

Halliday and Hasan (1976) explained that each information unit is structured in terms of two elements, a New element and a Given element (p. 326). While the Given element is optional, the New element is mandatory as without it there would be no information unit. Given information is that which the speaker or author presents as recoverable or accessible to the listener or reader via shared knowledge available either by previous

mention in the text or from outside the text. Halliday and Matthiessen (2014) proposed that while an "information unit does not correspond exactly to any other unit in grammar," it is nearest to the clause (p. 115). The preferred patterning in English is to present Given information first followed by New information, we can therefore hypothesize that the unmarked information status of a clause will be Given information followed by New information (Mahlberg, 2003, p. 104).

Mahlberg (2003) argued that lexical and phraseological patterns can also function to structure Given and New information. Mahlberg discussed general nouns, or nouns that capture a general meaning (e.g., *thing, man, fact, way*), positioning them as a device that can serve to introduce new information as they act as "a kind of hook onto which all the other information can be put" (p. 101). Mahlberg (2003) provided numerous examples:

1. … there's George Hamilton. *The man with chicken tikka complexion* pitches up in London this Saturday…
2. … what we got in America. *The thing is*, more of us are out of the…

In Example 1, *man* is the general noun that refers back to George Hamilton. Mahlberg explained that the general noun in this instance is used to add information in passing (p. 103). In Example 2, Mahlberg pointed out that the general noun *thing* provides "a kind of introduction or focusing device" to encapsulate Given information before introducing New information (p. 104). The larger phrases that the general nouns from Examples 1 and 2 are subsumed within, reflect the structure of frameworks: in Example 1, *the * with ** and in Example 2, *the * is.* The recurrent framework *the N1 of the N2* often houses general nouns, e.g., *the fact of the matter,* where general nouns and the frame combine to form a

lexical bundle that can act as a focusing device and hook off which to present new information. Sinclair (1991) also made the observation that in *of*-nominal groups the first noun will serve to highlight a specialized part, component, aspect, or attribute of the second noun (pp. 87-90). Examples include, *the first week of the war,* and *the blistering heat of the prairie.*

The next textual resource we will consider is that of connective. Quirk et al. (1985) explain that a "relation between parts of a text is achieved by connective features" (p. 1,437). Connectives often take the form of what Halliday and Hasan (1976) called "cohesive conjunctions" and Biber et al. (1999) called "linking adverbials." The two terms refer to cohesive devices that work to relate what follows the connective to the preceding text. Biber et al. (1999, pp. 875-879) provided a useful taxonomy for linking adverbials, dividing them into six meaning groups: (1) enumeration and addition (e.g., *for one thing, for another*), (2) summation (e.g., *in sum*), (3) apposition (e.g., *that is to say, for instance*), (4) result/inference (e.g., *as a result*), (5) contrast/concession (e.g., *on the other hand*), and (6) transition (e.g., *incidentally, by the way*). These meaning groups can help to pinpoint the textual function of connectives.

Reiteration is the last form of texture to be considered and has perhaps the clearest connection to simple lexis. Halliday and Hasan (1976) defined reiteration as one lexical item referring back to another (p. 278). These references could be the repetition of a word or related words, such as synonyms, hypernyms, or meronyms, throughout a text (see section 2.3). This reference of lexis creates a "lexical chain" that traverses a text, contributing to cohesion (see Figures 5 & 6, section 3.2.3, for visualizations of lexical chains).

1.3. Goals of the present study

The goal of the present study is to examine the use and role of the high frequency colligational framework *the N1 of the N2* in English argumentative essays authored by L1 speakers of English and Japanese. As developed throughout the introduction of this paper, the target framework is investigated with the following characteristics in mind: (1) semantic similarity of fillers in the first variable slot, and (2) the framework's role in creating texture via information structure, connectives, and reiteration. Specific research questions are:

1. How do the semantic characteristics of the fillers in the N1 position of the framework compare as used by L1 English and L1 Japanese authors of English argumentative essays?
2. What role does the target framework play in creating texture in English argumentative essays by L1 English and L1 Japanese authors?

2. Methodology

2.1. Corpus

The data used for this study were drawn from the written portion of the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2013). The ICNALE corpus consists of argumentative essays in response to two prompts:

> *Do you agree or disagree with the following statements? Use reasons and specific details to support your answer.*
>
> - Part-time job: It is important for college students to have a part-time job.
> - Smoking ban: Smoking should be completely banned at all the restaurants in the country.

These prompts were answered by groups of English language learners representing 10 different countries, as well as by L1 speakers of English from the United States, the United Kingdom, Australia, and New Zealand. The data used in the present study consisted of the 400 essays written by L1 speakers of English to make the ENS sub-corpus and a random sample of 400 essays from the 800 essays written by Japanese learners of English. Sub-corpora of equal sizes were desired for comparability. Word counts for the sub-corpora were generated via a custom script in the Python programming language and included words with apostrophes and hyphenated words as one word (e.g, *don't*; *part-time*) and also counted digits (e.g., 7). All statistical analyses were also carried out in the Python ecosystem using common data science packages (e.g., Pandas, NumPy, SciPy). Table 1 shows that the starting point for the data between the sub-corpora are similar in terms of overall raw word count and average length of .txt files. Type-token ratio, a measure of lexical diversity that divides the number of distinct word types by the total number of words (i.e., tokens), is also reported. A Welch's *t*-test revealed that mean type-token ratios of texts between L1 groups was significantly different at $p < .001$, with a large effect size $d = 1.48$. This indicates that the L1 English authors used a

Table 1. Overview of ENS and JPN sub-corpora used in the present study

	Sub-corpora	
	ENS	JPN
Word count	89,067	87,540
Number of texts	400	400
Mean Type Token Ratio	0.565 (0.05)*	0.492 (0.05)*
Mean text length	222.67 (23.85)*	218.85 (23.88)*

*Standard Deviation

broader array of vocabulary than the L1 Japanese.

2.2. Data extraction and storage

Following Eeg-Olofsson and Altenberg (1994), the present paper will consider instances of frameworks with one or two consecutive variable slots. Therefore, the target framework for the present study, *the N1 of the N2,* can include instances of modifiers preceding the nouns as well. In the event that the framework had modifiers preceding the main nouns, only the main noun was considered in the semantic analysis. The framework was extracted from the sub-corpora with a custom Python script that captured instances of one and two words in each variable slot. Frameworks were only included if they did not cross punctuation boundaries such as periods and commas. The extracted frameworks were stored in a SQLite database.

2.3. Semantic similarity of N1s

The semantic similarity of the N1s produced by the L1 English and Japanese writers was calculated using the WordNet lexical database (Miller, 2010). WordNet groups words into sets of synonyms called "synsets." The main relations that link words in synsets are hierarchical relations of synonymy, hypernymy, and meronymy. While the concept of synonymy is familiar to general audiences, hypernymy and meronymy might warrant further explanation. Hypernymy refers to relations between hypernyms and hyponyms, or superordinates and subordinates, respectively. An example of hypernymy is the word *car* in relation to *motor vehicle,* where *motor vehicle* is a hypernym of *car.* Hypernymy can be conceptualized as "isa" relationships as in a *car* ***isa*** *motor vehicle* (Hudson, 2007). Meanwhile, relations of meronymy are whole/part relations where the *whole* is the holonym and the *part* is the meronym. For example, if we talk about a

bumper as part of a *car*, the *bumper* is the part, or meronym, and the *car* is the whole or holonym. Relations of synonymy, hypernymy, and meronymy are depicted in Figure 1.

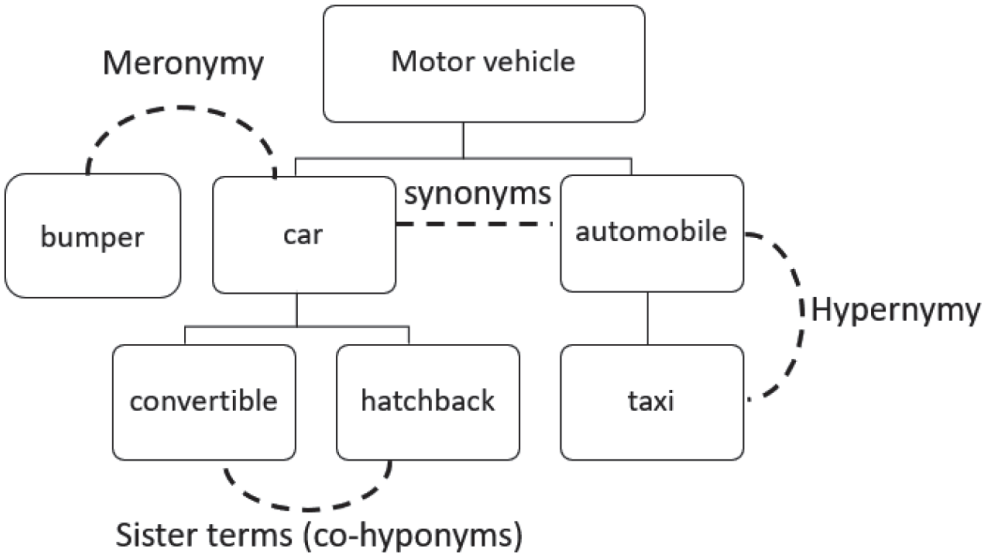

Figure 1. Schematic of semantic relations between words in WordNet

The basis of most measures of semantic similarity in WordNet is the distance between two words via connections of hypernymy, meronymy, or synonymy. For example, using Figure 1 as a reference, *hatchback* is only one hypernymic connection from *car,* but two from *motor vehicle.* Therefore, the similarity score between *hatchback* and *car* will be higher than that between *hatchback* and *motor vehicle.* Sun, Huang, and Liu (2011) provide an informative account of different WordNet-based measures of semantic similarity. They recommend the Wu-Palmer method as they feel it best aligns with human intuition (p. 123). Wu-Palmer scores of similarity range between 0 and 1 with the former indicating no semantic relation and the latter indicating synonymy. Figure 2 provides an example of Wu-Palmer scores and the lexical relation between the words that would give rise to

those scores. Note that the scores in Figure 2 will not apply to all words sharing the same relations in WordNet. That is, not all instances of direct hypernymy will have a Wu-Palmer score of .96 as seen in Figure 2. The reason is that scores are contingent on the number of layers in a hierarchy of words: the more layers, the smaller the decrease in scores between direct hypernyms.

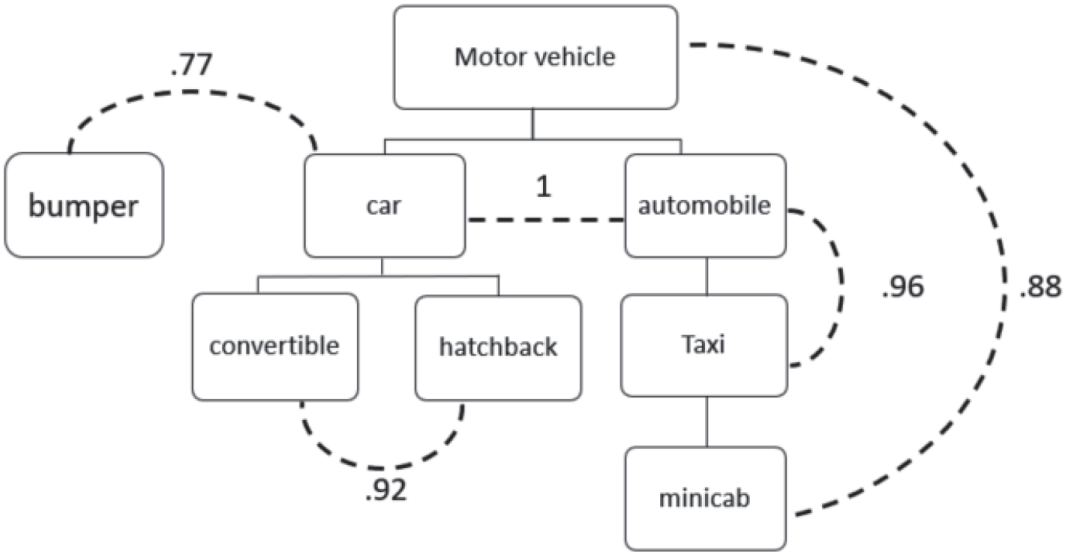

Figure 2. Example of semantic similarity scores between words in WordNet using Wu-Palmer

WordNet also disambiguates words for sense. For instance, the word *car* has five senses in WordNet. Sense 1 refers to the typical automobile most people might imagine when they hear the word *car*. Other senses include an *elevator car* and *cable car*. For this study, the author and a collaborator checked the sense of each N1 from the extracted frames and verified senses that were present in the corpus. Any disagreements were discussed and a final decision was negotiated. Therefore, similarity measures were only run between senses of words that were verified to be in

the corpus.

The calculation of Wu-Palmer similarity scores were automated in a Python script using the Natural Language Toolkit (NLTK) (Bird, Loper, & Klein, 2009). NLTK is a Python library for natural language processing. The comparison between two words with multiple senses that gave the highest Wu-Palmer score was kept to represent the relationship between the words. So, for example, two senses of the word *whole* appear in the N1 position of the target frame in the ENS sub-corpus and one sense of the word *middle*. Therefore, both senses of the word *whole* were compared with the one sense of the word *middle*. The resultant Wu-Palmer scores were 0.55 and 0.13, so only the higher score was kept to represent the semantic relationship between *whole* and *middle*. Once relationships between each N1 in the ENS sub-corpus and each N1 in the JPN sub-corpus were established, a basic network analysis was conducted to visualize the semantic relationship between the N1s using the network analysis package NetworkX for Python (Hagberg, Schult, & Swart, 2008).

2.4. The role of the frame in creating texture

The second analysis focuses on the framework's role in texture for each text. In particular, the three resources outlined in section 1.2 for creating texture are examined: (1) information structure, (2) connectives, and (3) reiteration. Only texts that contained an instance of the target feature were used: 71 and 45 files for the ENS and JPN sub-corpora, respectively. Given the manageable number of files, each file was manually inspected for the frame's role in creating texture. In order to count the types of cohesive ties, each frame was coded as carrying out a "primary function." If a frame clearly functioned to introduce new information and connect it back to Given information, it was counted as information structure. If the frame

clearly resided within a connective, such as a linking adverbial, it was counted as a connective.

Reiteration was operationalized in terms of repetition of a word lemma or a word related through synonymy, hypernymy, meronymy, or collocation. For repetition, Hoey's (1991) idea of complex repetition was adopted, meaning that derivational forms regardless of part of speech were considered repetition. For instance, *smoking, smoker,* and *smokes* were all considered repetition of the word *smoke*. Synonyms, antonyms, and words with relations of hypernymy and meronymy were identified via their relationships in WordNet. Collocation was operationalized using the word association measure Mutual Information (MI). MI values of 3.5 or higher in the Corpus of Contemporary American English (Davies, 2008), a more conservative measure than Hunston's (2002) suggestion of 3.0, were used.

3. Findings and Discussion

A brief profile of the overall make-up of files that contained the framework in each sub-corpus is given in Table 2. The tokens of the frame in the ENS sub-corpus exceed those of the JPN sub-corpus, 78 to 64, and are also more evenly dispersed among the texts occurring in over 90% of the ENS texts while occurring in only 70% of the JPN texts. This could suggest that L1 Japanese authors are not familiar with this common framework or are not comfortable using it and hence avoid it. It is also of note that while the framework is more frequent in ENS sub-corpus, the number of distinct N1s and N2s is fewer compared to the those in the JPN sub-corpus. This could be evidence of the L1 English authors drawing on a more restricted set of fillers reflecting better attunement to the colligational restrictions of the pattern.

Table 2. Data used for analysis from the ENS and JPN sub-corpora

	ENS	JPN
Number of instances of frame	78 (0.876)*	64 (0.731)*
Number of distinct fillers in the N1 position	42	46
Number of distinct fillers in the N2 position	44	54
Number of files the frame occurred in	71	45

*Normalized rate of occurrence per 1,000 words

3.1. Semantic characteristics of the N1 in the target frame

Table 3 shows all N1s that occur at least twice in either sub-corpus and their distribution between the smoking (SMK) and part-time job (PTJ) prompts. The italicized entries, *taste* and *health*, two abstract nouns, appeared in both the ENS and JPN sub-corpora and only in response to the prompt on smoking.

As Table 3 illustrates, the Japanese authors were more prompt-specific in their use of the N1s than their L1 English counterparts. In fact, there were no N1s in the JPN sub-corpus that occurred in texts representing both prompts. By contrast, nearly half of the N1s used by the L1 English authors appeared in texts in response to both prompts. This would seem to support the notion that the L1 English authors use more general, broadly applicable words in the N1 position than the L1 Japanese authors.

Semantic maps visually depicting the interconnectedness of the N1s used by the L1 English and Japanese authors are in Figures 3 and 4. The semantic maps intend to show how the N1s group together in a semantic space relative to their relationship to all the other N1s of the target frame from their respective corpora. Each N1 is represented by a circle, or node, in

Table 3. Proportions between prompts and N1s recurring at least twice in one sub-corpus.

	English	Freq (SMK:PTJ)	Japanese	Freq (SMK:PTJ)
1	fact	2 : 7	smoke	7 : 0
2	rest	6 : 3	smell	6 : 0
3	*health*	6 : 0	number	4 : 0
4	*taste*	5 : 0	*taste*	3 : 0
5	end	3 : 1	*health*	2 : 0
6	good	2 : 0	importance	0 : 2
7	effect	2 : 0		
8	interest	2 : 0		
9	majority	1 : 1		
10	owner	2 : 0		
11	part	2 : 0		
12	reality	1 : 1		
13	side	1 : 1		

*Words occurring in both lists in *italics*

the figure and are in color in the PDF version of this report. The nodes are connected by a green line to all other nodes that share a Wu-Palmer score of 0.25, or 25%, or higher. Node color reflects the number of connections a word shares with other words. The color transitions generally from dark blue to dark red as the number of connections increases. Blue nodes have the fewest connections in the map, and as the number of connections increases, the node color transitions to green, yellow, orange, and finally red signifying the most interconnected nodes in the map. Node size reflects the frequency with which an N1 appears in the framework: the more frequent the filler, the larger its node. So, for example, small blue nodes represent the lowest frequency words that share comparatively fewer semantic

connections to the other N1s in the map. Examples from the smaller cluster of nodes, Cluster 2, in Figure 3 (the ENS semantic map) are *development, side,* and *leader*. Large blue nodes represent words that filled the N1 position more frequently, but still feature comparatively fewer semantic connections to the other N1s. Large orange and red nodes represent N1s that are both comparatively more frequent and more interconnected with the other N1 fillers. Examples of frequent N1s that are also well interconnected in the ENS map are *fact, rest,* and *health,* all located in Cluster 1.

Interpreting the location of each node on the semantic maps is less transparent than color and size. To illustrate, we shall use specific examples from the ENS map. Note the nodes for *whole* and *effect* on the right side of Cluster 1. These two words share a high similarity score of 75%, which partially explains their proximity to each other. However, their similarity score is not the only factor that determines their position on the map. The position of each node is influenced by its relationship with all other nodes. Cluster 2 includes the words *development, owner, home, leader, middle,* and *side*. Judging by the lines extending from these words toward the main cluster, they all share connections to only *whole* and *effect* from Cluster 1. If we think of the lines connecting nodes as rubber bands pulling the nodes together, we can see why *whole* and *effect* are not more centralized in Cluster 1: because the blue nodes in Cluster 2 are working to pull them away. In general, the ENS semantic map shows that the N1s from the ENS sub-corpus appear to form one major semantic group in Cluster 1 of mostly well-connected nodes featuring words like *end, beginning, rest, majority,* and *future*. Additionally, the most frequent fillers in the N1 position are located in Cluster 1 (e.g., *fact, rest, health,* and *taste*).

The semantic map reflecting the relationship between the N1s in the

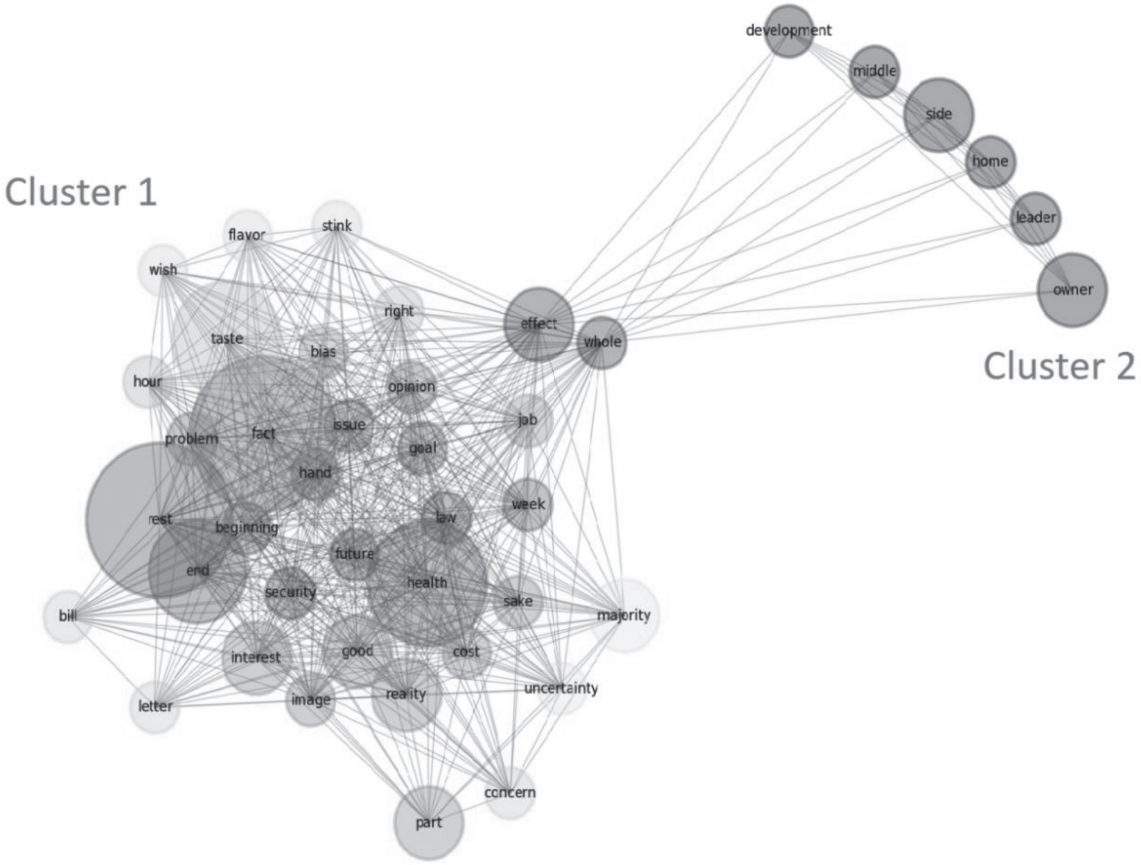

Figure 3. Semantic similarity map of N1s in *the N1 of the N2* framework in the ENS corpus (color is available on the PDF version of this manuscript available from the author upon request)

JPN sub-corpus displayed in Figure 4 is quite different from that of the ENS sub-corpus. Some striking differences between the ENS and JPN maps are that (1) unlike in the ENS map, there is an overall dearth of orange or red nodes in the JPN map, meaning that the bulk of the N1s are comparatively less interconnected than the central nodes *balance* and *cause*; (2) the most frequent N1s in the JPN map are not exclusively found in one coherent semantic group as we saw with the N1s in the ENS map; and (3) while the fillers are separated into two distinct clusters in both figures, the ENS map sees the majority of N1s concentrated in a single cluster. In the JPN map, the highly-connected words *balance* and *cause*

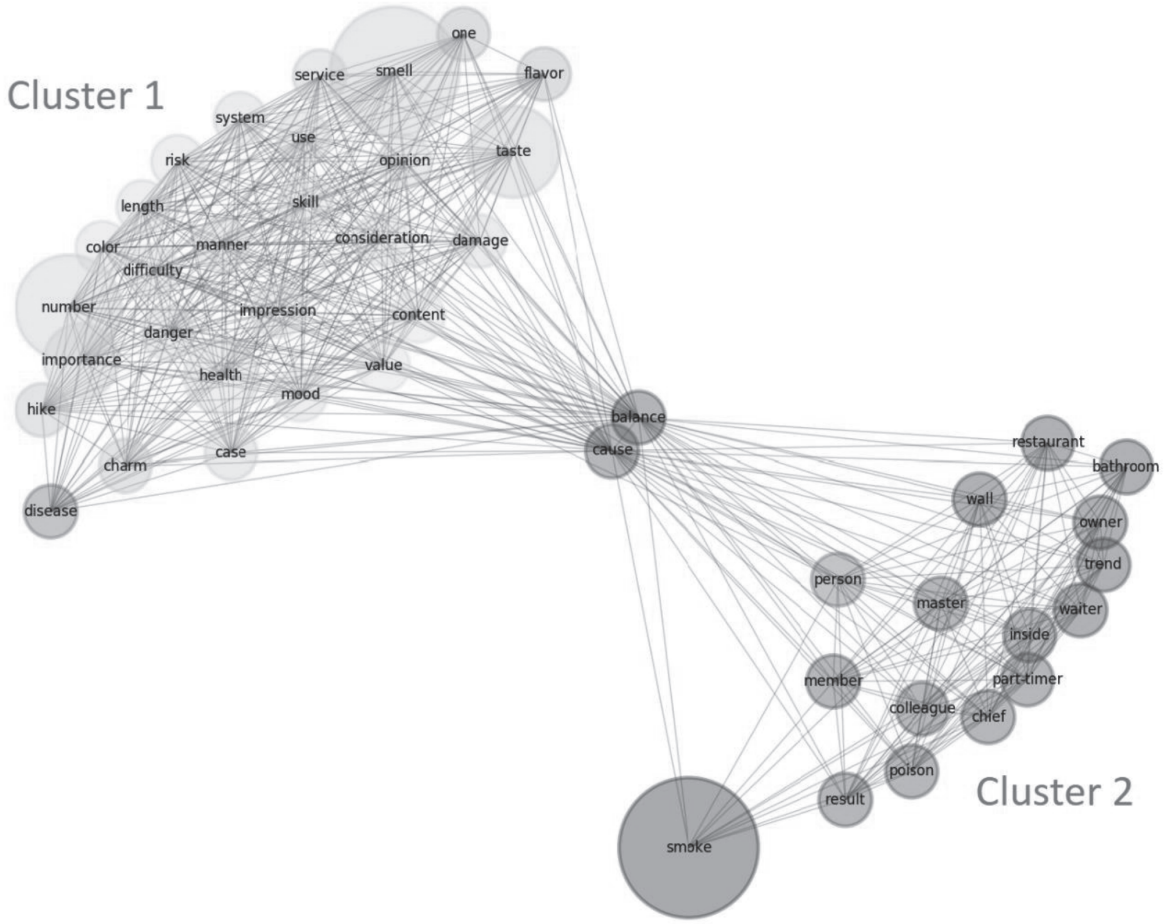

Figure 4. Semantic similarity map of N1s in *the N1 of the N2* framework in the JPN corpus (color is available on the PDF version of this manuscript available from the author upon request)

hover between Clusters 1 and 2 which are much closer in size than the two clusters in the ENS map. In essence, the Japanese authors are using two distinct semantic groups of fillers that are closer in size, and the two words that bridge those groups, *balance* and *cause,* are pulled to a semantic space that resides more or less evenly between them.

Perhaps the most telling observation of the groupings of N1s in Figures 3 and 4 is the types of nouns in Clusters 1 and 2 of the respective figures. Cluster 1 of both figures features words that are all abstract entities in WordNet's hierarchy of hypernyms. To provide a more precise

categorization than “abstract entity,” each word was classified using Biber’s (2006) semantic categories for nouns (pp. 248-250). Drawing on nouns that occurred more than 20 times per million words in the T2K-SWAL Corpus, Biber (2006, p. 248) identified eight categories of nouns reproduced here:

Animate: humans or animals

Cognitive: mental/cognitive processes or perceptions

Concrete: inanimate objects that can be touched

Technical/concrete: tangible objects that are not normally perceived and/or cannot normally be touched

Place: places, areas, or objects in a fixed location

Quantity: nouns specifying a quantity, amount, or duration

Group/institution: nouns that denote a group or institution

Abstract/process: intangible, abstract concepts or processes

Table 4 presents the semantic category of each noun in the N1 position from the ENS sub-corpus in terms of Biber’s (2006) semantic categories for nouns based on their cluster membership in Figure 3. The first column of Table 4 presents the semantic category, the second column presents all N1s located in Cluster 1, and the third column presents all N1s located in Cluster 2. The differences in types of nouns that make up the clusters is striking. For instance, 21 of the 23 abstract/process nouns that make up the ENS map belong to Cluster 1. All eight of the cognitive nouns belong to Cluster 1, and the remaining seven nouns of Cluster 1 are divided between quantity and technical concrete nouns. Cluster 1 contains no animate, place, or group/institution nouns while Cluster 2 does feature nouns representing each of those categories. Clearly, the vast majority of N1s used by the L1 English authors are abstract/process and cognitive nouns.

Table 4. Semantic categories of N1s in ENS sub-corpus

Semantic category	Nouns in Cluster 1	Nouns in Cluster 2
Animate		*owner, leader*
Cognitive	*concern, fact, flavor, opinion, stink, taste, uncertainty, wish*	
Concrete		
Technical concrete	*bill, letter*	
Place		*middle*
Quantity	*future, hour, majority, part, rest, week*	
Group/institution		*home*
Abstract/process	*beginning, bias, cost, effect, end, goal, good, hand, health, image, interest, issue, job, law, problem, reality, right, sake, security, whole*	*side, development*

Table 5 presents the semantic category of each N1 from the JPN sub-corpus in. The fillers *balance* and *cause* are included with Cluster 1 as they share more semantic ties with Cluster 1 than with Cluster 2.

Similarities between the types of nouns that fill the N1 position in the ENS and JPN sub-corpora can be seen in Tables 4 and 5, and are reflected in how the N1s clustered in the semantic maps. Cluster 1 in both maps features more interconnected words and contains mostly abstract/process and cognitive nouns such as *taste, health, wish, end, rest,* and *development*. These types of nouns define some intangible aspect of the N2 as illustrated in the following concordances from the JPN and ENS sub-corpora:

3. …the smoke ruins *the taste of the dish* and the smoking with… (JPN, SMK)

Table 5. Semantic categorization of N1s in JPN sub-corpus

Semantic categories	Nouns in cluster 1	Nouns in cluster 2
Animate		*chief, colleague, master, member, owner, part-timer, person, waiter*
Cognitive	*consideration, flavor, impression, mood, opinion, smell, taste*	
Concrete		*poison, smoke, wall*
Technical concrete	*disease*	
Place		*bathroom, inside, restaurant*
Quantity	*length, number, one*	
Group/institution		
Abstract/process	*balance, case, cause, charm, color, content, damage, danger, difficulty, health, hike, importance, manner, risk, service, skill, system, use, value*	*result, trend*

4. … smoking has a risk to *the health of the people* who have been troubled… (JPN, SMK)
5. … governments would be following *the wishes of the majority* of cities. (ENS, SMK)
6. When it comes down to it, at *the end of the day* that's really all that… (ENS, SMK)
7. … in the summers and part time through *the rest of the school year*… (ENS, PTJ)

While there is much overlap in the types of N1s that comprise the ENS and JPN semantic maps, there are also differences. Table 6 compares the

proportion of N1 semantic types used by L1 English and Japanese authors. Abstract/process nouns made up 45% of the tokens used in the N1 position by the L1 English authors compared to 36% by the L1 Japanese authors. Quantity nouns also made up much more of the N1s in the ENS corpus accounting for 21% of the tokens compared to 9% in the JPN sub-corpus. Outside of abstract/process and quantity nouns, the biggest differences in types of nouns used by the two groups of writers was concrete nouns and animate nouns. The L1 English authors used no concrete nouns in the N1 position compared to 14% of the fillers from the JPN corpus, and animate nouns made up only 4% of the fillers in the ENS corpus but 13% of the JPN corpus.

Table 6. Percent of N1 types in frames out of N1 tokens

Semantic categories for nouns (Biber, 2006)	ENS	JPN
Animate	4%	13%
Cognitive	26%	22%
Concrete	0%	14%
Technical concrete	3%	2%
Place	5%	5%
Quantity	21%	9%
Group/institution	1%	0%
Abstract/process	45%	36%

What most clearly separates the two groups' use of the frame is the Japanese authors' more frequent use of the genitive *of*-phrase with animate and concrete nouns and the L1 English authors' use of the more idiomatic phrases such as the *fact of the matter* and *(at) the end of the day*. The more

idiomatic phrases will be treated in more depth in the next section on the frame's role in creating texture in the text. With respect to the genitive *of*-phrase, the Japanese authors used it frequently with *smoke* in the N1 position. *Smoke* is a concrete noun. Recall that the L1 English authors did not use concrete nouns in the N1 position of the frame in this data set. Rather, they opted for the phrase *cigarette smoke* 89 times.

8. …*the smoke of the cigarette* is worse for a woman and a minor (JPN, SMK)
9. …it is considered that *the smoke of the cigarette* is bad for not only… (JPN, SMK)
10. …you do not have to smell *cigarette smoke* which interferes with the… (ENS, SMK)
11. …enjoying food, and having *cigarette smoke* mixed into the air… (ENS, SMK)

There was also more use of animate nouns in the frame in the JPN sub-corpus that were followed by *of*-phrases functioning as a post noun modifier. Take, for example, the use of *colleagues* and *waiters*:

12. …we can make friends with *the colleagues of the part-time job*. (JPN, PTJ)
13. …as a visitor, but *the waiters of the restaurant* must come to work (JPN, SMK)

In Examples 12 and 13, the N1 already carries meaning that the *of*-phrase brings, *colleagues* are at work and *waiters* work in restaurants, rendering the information that the prepositional phrase carries redundant.

In summary, the Wu-Palmer method of calculating similarity did a reasonable job separating out nouns of different semantic categories, and

placed them relative to one another in a two-dimensional space. The findings indicate that particularly for the L1 English authors, the majority of N1s belong to a semantically interconnected group of abstract/process and cognitive nouns. The N1 also typically shares an interesting relationship with the N2 in that it tends to define some abstract or intangible quality or attribute of the N2. This latter finding is congruent with Hasselgård's (2019) findings on the same framework. The next section will investigate how the framework contributes to creating texture in the texts.

3.2. The role of the frame in creating texture

In section 1.2 three manners in which formulaic language can contribute to texture were outlined: (1) information structure, (2) connectives, and (3) reiteration. The following sub-sections will look more closely at how the frame was used to create texture.

3.2.1. Framework in Information Structure

Use of information structure as a cohesive resource was slightly more frequent in the JPN texts than the ENS texts at 0.22 and 0.16 times per 1,000 words, respectively. In the JPN corpus, the frame frequently holds the subject position of the clause: *the smoke of the cigarette, the smell of the cigarette, the smell of the smoke.* These instances hold a noun phrase and represent Given information that leads into a new proposition as shown in the following examples:

14. Thirdly, *the smoke of the cigarette* makes the room dirty. (JPN, SMK)
15. *The smell of the cigarette* is never comfortable for non-smokers. (JPN, SMK)
16. In addition, I think that *the smell of the smoke* ruins the taste of the

dish… (JPN, SMK)

However, there are some differences in how information structure is used in the ENS and JPN sub-corpora. With the JPN authors, over half of the 19 instances of the frame contain Given information in both the N1 and the N2. Also, the N2 is most frequently a concrete or place noun such as *cigarette, people,* or *restaurant* and used in the genitive *of*-phrase. Meanwhile, in the ENS sub-corpus, more abstract nouns appear in the N1 position. Recall Mahlberg's (2003) example of a focusing device with the phrase *the thing is,* where *thing* is a general noun used to encapsulate and focus Given information before presenting New information. The most frequent example from the ENS sub-corpus is *the fact of the matter*. Like the JPN texts, the framework houses a noun phrase in the subject position of the clause. Unlike the examples from the JPN sub-corpus, it uses a general abstract noun, *matter*, to package the entirety of the Given information. This is seen in Example 17. Example 18 uses the noun *many* in the N2 position to refer back to the general population of Japan and how its well-being should concern individuals in Japan.

17. *The fact of the matter* is that we are already busy enough… (ENS, PTJ)
18. … where *the good of the many* is the concern of the individual (ENS, SMK)

3.2.2. Framework as Connective

With only 6 and 4 raw instances in the ENS and JPN sub-corpora, respectively, the framework did not frequently function as a connective. When the framework did function as a connective in the ENS corpus, it was as a linking adverbial and frequently an idiomatic expression. For example:

19. At *the end of the day*, a student's most important directive is to go to school (ENS, PTJ)
20. But at *the end of the day*, I don't really know. (ENS, SMK)
21. On *the other side of the coin*, I know that a lot of students are fairly resilient and… (ENS, PTJ)

Using Biber et al.'s (1999) semantic categories of linking adverbials, Example 19 fulfills the function of "summation" in that the author used it to "conclude or sum up the information in the preceding discourse" (p. 876). Example 20 also functions to sum up the author's opinion on the topic of the text. However, the author used two linking adverbials in succession with "But" functioning as "contrast/concession" to precede his or her ultimate conclusion. The author points out the health drawbacks of smoking as support for the thesis of banning smoking in public places, but the author concedes that the choice ultimately rests with the Japanese people. Example 21 also functions for "contrast/concession" by linking an opposing view to the preceding text.

The connectives in the JPN sub-corpus were less formulaic and idiomatic than the connectives in the ENS sub-corpus. In Examples 22 and 23, the connectives work to link what follows to the preceding discourse, but the choice of words differs from what an L1 or more proficient speaker would typically use. In Example 22, a more proficient speaker would most likely use *as a result of the above*, and in Example 23 most likely *in the worst-case scenario*.

22. As *the result of the above*, protecting non-smokers from side-stream smoke is the most important thing… (JPN, SMK)
23. …and, in *the case of the worst*, miscarriage and to have a baby born… (JPN, SMK)

While the frame does not result in many connectives for either group of writers, the L1 English writers tended to use idiomatic sequences of words to achieve texture while the Japanese authors appear to be less aware of the idiomatic phrases and instead pieced together individual words resulting in unnatural sequences.

3.2.3. Framework in Reiteration

Previous research on lexical overlap and writing development over time has found that there is an increase in cohesive devices such as reiteration and connectives as young writers initially develop, but this trend reverses as writers mature and improve over time (Crossley et al. 2011; Haswell, 1990). Much like young L1 learners, L2 learners also seem initially to use more explicit cohesive devices before the trend reverses later in development (Crossley et al. 2016; Yang & Sun, 2012). In light of these trends, it was expected that the N2 would feature more reiteration throughout the texts in the JPN sub-corpus due to a typically narrower range of vocabulary when compared to L1 English authors.

The N2 was chosen to be the focus of this part of the analysis because it is usually the N2 of this framework that serves as the head noun (Sinclair, 1991). Figures 5 and 6 are examples of reiteration in a text from the ENS and JPN sub-corpora, respectively. The text samples feature the target frames, ***the middle of the restaurant*** and ***the mood of the restaurant***, marked in ***bold italics***, and the N2 is also ***underlined***. Instances of reiteration of the N2 throughout the texts via complex repetition are marked by plain **bold text** and instances via collocation by plain underlined text. There were no instances of reiteration via synonymy, hypernymy, or meronymy in these examples. Lines to delineate the lexical chains that traverse the texts are provided.

In an essay from the ENS sub-corpus presented in Figure 5, it is seen that the author begins the essay with the topic sentence about "Banning smoking at restaurants…". The author uses the frame to emphasize how inconsiderate it is to smoke in the restaurant by pinpointing a specific attribute of the restaurant, "the middle," as the location where the patron chooses to "light up." The N2, *restaurant,* has five anaphoric and cataphoric links creating a lexical chain anchored at more or less equal intervals throughout the text: four instances of repetition and one instance of collocational overlap with the word *meal*.

Banning smoking at **restaurants** is becoming a common practice in many states in the United States of America. The fact that smoking is still permitted at Japanese **restaurants** would thus appear to be a temporary condition, because as time goes on the limitations on smoking are becoming more and more widespread. In addition, no one likes to have their meal ruined by an inconsiderate individual who deems it appropriate to light up right in ***the middle of the restaurant***. Such an individual is endangering not only himself but also others around him with his second hand smoke, which is surprisingly much more dangerous than the smoke he will be inhaling himself. Long periods of exposure to secondhand smoke have been known to cause cancer just as long periods of smoking does as well. With this kind of obvious link between a life threatening disease and a stupid habit, the choice to quit smoking and the choice to ban smoking at **restaurants** would seem to be obviously a good one. Nonetheless, there are many individuals who consider themselves rebels and find all kinds of ways to claim that other people are imposing upon their rights to be themselves. There is no way to please this kind of individual, but there is a way to keep this kind of individual from hurting other people. That would be by banning smoking at the **restaurants** in Japan.

Figure 5. Lexical cohesion with the N2 *restaurant* in a text from the ENS sub-corpus

Figure 6 depicts a text from the JPN sub-corpus and follows the lexical chain for the N2 of the frame *the mood of the restaurant*. The relationship between the N1 and the N2 in this instance lines up well with that in the ENS corpus as *mood* defines some intangible aspect of *the restaurant*. Also, the frame identifies *mood* as an important attribute of the *restaurant*. The

more pronounced difference between Figures 5 and 6 is the number of semantic connections to the N2 throughout the texts. While, the text from the ENS sub-corpus features five semantic connections to the N2, the text from the JPN sub-corpus features eight connections.

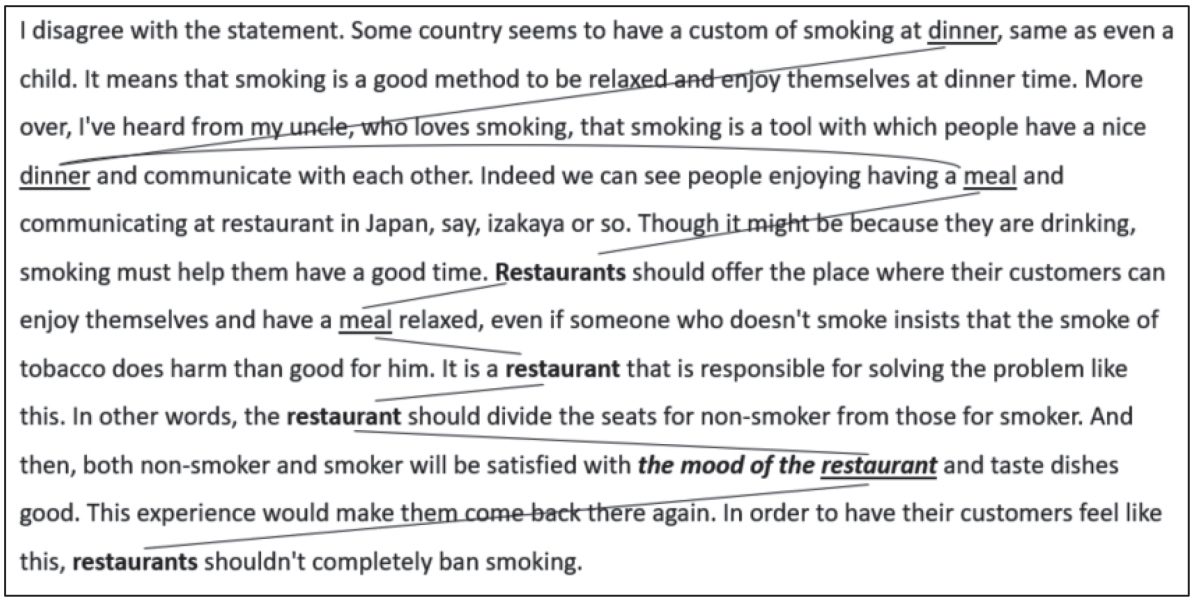
I disagree with the statement. Some country seems to have a custom of smoking at dinner, same as even a child. It means that smoking is a good method to be relaxed and enjoy themselves at dinner time. More over, I've heard from my uncle, who loves smoking, that smoking is a tool with which people have a nice dinner and communicate with each other. Indeed we can see people enjoying having a meal and communicating at restaurant in Japan, say, izakaya or so. Though it might be because they are drinking, smoking must help them have a good time. **Restaurants** should offer the place where their customers can enjoy themselves and have a meal relaxed, even if someone who doesn't smoke insists that the smoke of tobacco does harm than good for him. It is a **restaurant** that is responsible for solving the problem like this. In other words, the **restaurant** should divide the seats for non-smoker from those for smoker. And then, both non-smoker and smoker will be satisfied with ***the mood of the restaurant*** and taste dishes good. This experience would make them come back there again. In order to have their customers feel like this, **restaurants** shouldn't completely ban smoking.

Figure 6. Lexical cohesion with the N2 *restaurant* in a text from the JPN sub-corpus

As outlined in Table 7, the mean number of total links an N2 had throughout the texts in the ENS sub-corpus was 3.99—less than half the number of links that texts in the JPN sub-corpus featured with a mean of 9.00. These numbers reflect all lexical links such as lexical overlap of the lemma, synonyms, antonyms, collocates, hypernyms, and meronyms

Table 7. Means, standard deviations, and range of lexical links to N2 per text

	Mean	SD	Range
ENS	3.99	3.56	0-17
JPN	9.00	7.90	0-32

appearing elsewhere in the text. It is clear that, in general, there are more instances of reiteration related to the N2 in the essays from the JPN sub-corpus.

A more detailed look at the types of reiteration that make up the lexical chains in the ENS and JPN sub-corpora is provided in Table 8. Texts from the JPN sub-corpus have an average of more than 3.0 instances per text for both repetition of the lemma and collocational overlap, while the ENS texts have a comparatively lower rate of occurrence at approximately 1.6 per text for both categories. The average per text rate between groups is much more similar in the measures of synonymy, antonymy, hypernymy, and meronymy. Higher rates of repetition among the L1 Japanese authors is likely related to the observation that identical lexical overlap is more common in both early first language and second language learner writing but does little in the way of elaborating ideas in a text and hence decreases as writers develop (Haswell, 1990). This may also be related to less lexical diversity of the texts in the JPN sub-corpus compared to those in the ENS sub-corpus as was reflected in the type-token ratios. That is, the Japanese authors have fewer words at their disposal in English and are therefore more likely to repeat words, inflating rates of repetition.

Table 8. Mean per text frequency of types of lexical links with the N2 in the ENS and JPN corpora

	Repetition	Collocation	Synonymy	Antonymy	Hypernymy	Meronymy
ENS	1.67	1.61	0.20	0.03	0.62	0.14
JPN	3.27	3.04	0.24	0.02	0.71	0.09

4. Conclusion

This study contributed empirical evidence supporting the intuition that N1s in the colligational framework *the N1 of the N2* share semantic characteristics. By using network analysis, it was shown that in this data set most N1s share a semantic space defined by hypernymic and meronymic relationships. Furthermore, the L1 English authors appear to choose N1s from a more restricted semantic space than their L1 Japanese counterparts. In terms of the relationship between the N1 and the N2, the L1 English authors typically fill the N1 slot with an abstract/process or cognitive noun that defines some intangible aspect or attribute of the N2; this is congruent with Hasselgård's (2016 & 2019) characterization of the semantic relationship between the N1 and N2 in this framework. The Japanese authors do this to a lesser extent, relying on more animate and concrete nouns in the N1 position, particularly in the genitive *of*-phrase.

In terms of creating texture, the framework contributed to information structure, connectives, and reiteration. The L1 English authors used the framework for information structure less than the L1 Japanese authors, and the nature of the usage also differed between groups. The L1 English authors more frequently used general nouns in idiomatic expressions to focus the preceding text in its entirety before presenting New information than the L1 Japanese authors. In terms of connectives, the L1 English authors used more idiomatic linking adverbials than the Japanese authors whose use of connectives in the framework appeared to be constructed piecemeal. With respect to reiteration, the vast majority of N2s in the framework represented Given information featuring some sort of reiteration of previously established entities. However, the nature of the textual links differed between the two groups. The Japanese authors used

more instances of repetition and collocation with the N2 than the L1 English authors, but both groups feature similar frequencies of links established through other forms of semantic repetition such as synonymy and hypernymy.

These findings have important implications for the teaching of this high-frequency framework as it plays an important role in text construction and cohesion. One immediate takeaway in the context of Japan is to raise awareness about differences in usage of the frame between the groups of authors. Instructors could create data-driven learning activities that focus on induction of patterns and exemplar-based learning to guide learners to observe these differences (Boulton & Cobb, 2017). For example, instructors might have learners compare instances of the frame from the ENS and JPN sub-corpora to try to discover the qualitative variation between the fillers in the N1 position (e.g., abstract versus concrete nouns). This could lead to a discussion about the differences between abstract and concrete nouns and how the frame can be used to highlight some intangible aspect of an entity versus the genitive *of*-phrase. Learners could then re-examine the samples of language in light of the class discussion. The goal would be to guide learners to reconsider how and when they want to use this high-frequency frame.

Likewise, activities that guide learners to contemplate the relationship of the framework to the larger text could be useful. It is apparent that texts in the JPN sub-corpus featured more instances of creating cohesion via lexical repetition with the N2 of the framework. While this is likely an artefact of a more restricted English lexicon, it could raise awareness of the trend and serve to motivate learners to aim for more lexical breadth in their writing. Furthermore, noticing the more idiomatic instances of the framework (e.g., *the fact of the matter* and *(at) the end of the day*) that

function as part of information structure and connectives could raise awareness of norms for structuring information flow.

Of course, there are numerous limitations to this study. For one, the sample of texts was small, and the subset of texts that had the target feature was even smaller. Cortes (2015, p. 205) explained that problems can arise when making comparisons with or between small corpora. Future studies might amend these weaknesses by re-visiting the topic with larger corpora. Another avenue to pursue would be to examine the role of discontinuous formulaic language with a wider range of different first language groups and in different registers. Finally, alternative methods of gauging semantic similarity such as distributional methods that have become more commonplace in computational linguistics (e.g., word2vec) would be a welcome extension.

Language is a complex phenomenon and the empirical findings here point to subtle differences in the usage of a high frequency instance of formulaic language between more and less proficient speakers of English. Furthermore, the differences in the usage of the framework found between L1 speakers and learners may warrant more critical contrastive interlanguage analysis over a wider scope of discontinuous formulaic language. Such analyses have the potential to guide instructors toward innovative teaching materials to address the differences between L1 groups outlined in the present article.

References

Biber, D. (2006). University language: A corpus-based study of spoken and written registers. Amsterdam: John Benjamins.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics, 14*(3), 275-311. doi: 10.1075/ijcl.14.3.08bib

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405. doi: 10.1093/applin/25.3.371

Biber, D., & Gray, B. (2016). *Grammatical Complexity in Academic English: Linguistic Change in Writing.* Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman Grammar of Spoken and Written English.* London: Longman.

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python.* Cambridge: O'Reilly Media Inc.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning, 67*(2), 348-393. doi: 10.1111/lang.12224

Cortes, V. (2015). Situating lexical bundles in the formulaic language spectrum: origin and functional analysis developments. In V. Cortes, E. Csomay, & D. Biber (Eds.), *Corpus-based Research in Applied Linguistics: Studies in Honor of Doug Biber* (pp. 197-216). Amsterdam: John Benjamins. Council of Europe. 2001.

Crossley, S. A., Roscoe, R. D., McNamara, D. S., & Graesser, A. (2011). Predicting human scores of essay quality using computational indices of linguistic and textual features. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th international conference on artificial intelligence in education* (pp. 438–440). New York: Springer.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *The Journal of Second Language Writing, 32,* 1-16, doi: 10.1016/j.jslw.2016.01.003

Davies, M. (2008) The Corpus of Contemporary American English: 1 billion words, 1990-present. http://corpus.byu.edu/coca/

Eeg-Olofsson, M., & Altenberg, B. (1994). Discontinuous recurrent words combinations in the London-Lund Corpus. In U. Fries, G. Tottie, & P. Schneider (Eds.), *Creating and Using English Language Corpora: Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora.* (pp. 63-77). Amsterdam: Rodopi.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis,* Oxford: Blackwell, 1-32.

Garner, J. R. (2016). A phrase-frame approach to investigating phraseology in learner writing across proficiency levels. *International Journal of Learner Corpus Research,* 2(1),31-67. doi: 10.1075/ijlcr.2.1.02gar

Gray, B., & Biber, B. (2015). Lexical frames in academic prose and conversation. In

S. Hoffmann, B. Fischer-Starcke, & A. Sand (eds.), *Current Issues in Phraseology* (pp. 109-133). Amsterdam: John Benjamins.

Hagberg, A. A., Schult, D. A., & Swart P. J. (2008). Exploring network structure, dynamics, and function using NetworkX, in *Proceedings of the 7th Python in Science Conference (SciPy2008),* Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008.

Halliday, M. A. K., & Hasan. R. (1976). *Cohesion in English*. London: Longman.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's introduction to functional grammar* (4th ed.). London and New York: Routledge.

Hasselgård, H. (2016). "*The way of the world*: The colligational framework 'the N1 of theN2' and its Norwegian correspondences". *Nordic Journal of English Studies*, *15*(3):55–79. doi: http://ojs.ub.gu.se/ojs/index.php/njes/article/view/3589

Hasselgård, H. (2019). The nature of the essays: The colligational framework 'the N of the N'in L1 and L2 novice academic English. *Corpus approaches into World Englishes and Language Contrasts* (Studies in Variation, Contacts and Change in English 20), ed. by Hanna Parviainen, Mark Kaunisto & Päivi Pahta. Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/20/hasselgard/

Haswell, R. H. (1990). *Change in undergraduate and post-graduate writing performance (part 2): Problems in interpretation (ERIC No. ED323537)*. ERIC Institute of Education Sciences. Retrieved from (http://eric.ed.gov/?id=ED323537)

Hoey, M. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press.

Hudson, R. (2007). *Language networks: The new word grammar*. New York: Oxford University Press.

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

Hyland, K. (2008). As can be seen: Lexical bundles and discplinary variation. *English for Specific Purposes*, 27, 4–21. Doi: 10.1016/j.esp.2007.06.001

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* (pp. 91-118). Kobe: Kobe University School of Languages and Communication.

Mahlberg, M. (2003). The textual dimension of corpus linguistics: The support function of English general nouns and its theoretical implications. *International Journal of Corpus Linguistics, 8*(1), 97-108.

Marco, M. J. L. (2000). Collocational frameworks in medical research papers: a genre-based study. *English for Specific Purposes*, *19*(1), 63–86.

Miller, G. A. (2010). "About WordNet." WordNet. Princeton University. Retrieved from http://wordnet.princeton.edu

Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signaling in academic lectures. *International Journal of Corpus Linguistics, 11*(3), 283-304. doi: 10.1075/ijcl.11.3.04nes

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive Grammar of the English Language*. London & New York: Longman.

Renouf, A., & Sinclair, J. McH. (1991). Collocational frameworks in English. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics*. London: Longman, 128–143.

Römer, U. (2010). Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction*, *3*(1), 95–119. doi: 10.1075/etc.3.1.06rom

Sinclair, J. McH. (1991). Corpus, Concordance, and Collocation. Oxford: Oxford University Press.

Sun, K., Huang, Y., & Liu, M. (2011). A WordNet-Based Near-Synonyms and Similar-Looking Word Learning System. *Educational Technology & Society*, *14*, 121–134.

Yang, W., & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education, 23*, 31-48.