Universe Scientific Publishing

ORIGINAL ARTICLE

An N-gram Analysis Japanese English Learners' Writing: Focus on Grammatical Patterns

Lingzhi Wang

Busan University of Foreign Studies, Busan 46234, Korea.

Abstract: The purpose of the study is to investigate how Japanese English learners use formulaic sequence of language in their writing. Formulaic sequences, for example, n-grams or chunks are considered to be one of the criteria for teachers to level learners' language naturalness. The study analyzes Japanese and native learners' academic writing, and finds that Japanese learners always over-use/under-use certain n-grams. This indicates that Japanese learners prefer to use n-grams that is not frequent used by native speakers. At the end of the study, some teaching methods and implications will be given according to the results.

Keywords: N-gram; Grammar Patterns; Japanese English Learners; Academic Writing

Hoey (2004) points out that continuous sequences of words usage are considered to crucial features that can suggest native-like language use. The theoretical background to this issue is that native speakersrely more on prefabricated word sequences in their language use (Adel & Erman, 2012). This formulaic sequences or "chunks" constitute the key element of fluent processing. Moreover, there is a trend that native speakers are likely to use more fixed sequences of words in spoken language than in written language (Pawley & Syder, 1983). However, native speakers tend to have a unique way of using formulaic language in particular style. To English as first language (EFL) or English as second language(ESL), this particular style of fixed phrases means "naturalness" or "normality" of native English use. Therefore, EFL/ESL teachers should have enough ability and understanding of formulaic language mechanism so that students can be facilitated with their help when they have difficulties in appropriate collocations.

With the development of corpus linguistics, more and more researchers and scholars are encouraged to explore collocational patterns of continuous word sequences in second language acquisition area (Biber, Johansson, Leech, Conrad & Finegan, 1999; Hong, 2012). In addition, the WordSmith Tools is quite easy for researchers to extract n-gram from corpus using user-friendly concordancers (e.g. WordSmith 4.0). N-gram refers to a continuous sequence of words, such as according to (2-gram), a lot of (3-gram), or at the end of (4-gram). N-gram analysis has provided a useful method and promoted teachers' understanding of distinctive patterns, however, although corpus-based chunks can provide a beneficial and effective syllabus design and approaches, it is essential to pedagogically useful n-grams so that learners can achieve a best effect in language naturalness (Hong, 2019). On this account, this study will focus on exploring how Japanese learners use 4-gram in terms of grammar patterns and seek to give some implications to teaching the language naturalness of collocations for Japanese learners.

1. Literature review

Recently, a great deal of studies concerning formulaic patterns is based on corpus approach, however, there is one problem appears in the aspect, which different terminologies of analyzing these kinds of patterns spring up (Wray, 2002). An identified definition of n-gram is necessary, since the study need to use n-gram analyzing clearly. Ten different terminologies that commonly used from a corpus-based approach are distinguished by Hong (2013). According to Hong (2012) these ten terminologies can be divided into three groups. Group 1 includes phraseology, formulaic sequence, phrasicon, concgram, which share with the definition of "co-occurring word patterns (not necessarily continuous)". Group 2 is constituted of lexical bundle, n-gram, multi-word construction, cluster, recurrent word combination with a shared definition of "repeated continuous

doi: 10.18686/ahe.v5i3.3474

sequences of words". Group 3 has only one skipgram in it, which means that "non-continuous sequence of words using a skip distance of n. N-gram used here in this study with Hong's (2013) definition "an n-gram refers to a continuous sequence of n-words" is because n-gram is initially used as a computer terminology and the continuous sequence of words is automatically extract by n-gram algorithm in WordSmith without grammatical or syntactic considering. For this reason, the researcher prefers to use n-gram rather than other terminologies.

The key notion of n-gram is originated from Sinclair's (1991) theory of idiom principle. He claims that words occur as a form of semi-preconstructed phrases which have primacy over grammar, the open-choice principle in his terminology.

These theories of formulaic words are confirmed by corpus linguistics which offers useful methods and techniques with effective evidences of recurrent patterns (McEnery and Hardie, 2012). One noticeable matter is that many re-appeared sequences are non-compositional or idiomatic with little chance to substitution (Simpson-Vlach and Mendis, 2003). Moreover, native speakers may tend to use a particular expression rather than alternatives since they have their own conventionalized language usage in their community. Since n-grams are retrieved by an automatic concordance with specific cut-off points, they are like idiomatic expressions. A key point here is that the non-compositional features of an n-gram is a surface presentation of frequency list generated by a concordance. Therefore, n-gram studies in terms of grammatical and functional patterns should give a standard frequency cut-off point (Biber *et al*., 1999).

Hong (2013) investigate EFL Korean learners' use of continuous words sequences. He compiles a small corpus of Korean learners' essays for contrastive analysis with a native learners' corpus, and extracts lists of 4-word sequences (4-grams) form the Korean learners corpus and the native learners corpus. The results indicate that Korean learners use 4-grams differently compare to native speakers. It shows an overuse of NV (Noun-Verb phrase) among Korean learners, unlike native speakers who use PP (prepositional Phrase) more.

2. Method

2.1 Research question

Do Japanese learners have the same use of formulaic patterns to native speakers?

2.2 Data collection and participants

The study analyses the academic essays of Japanese English learners and native speakers. 366 undergraduate students majoring in English language and literature participate in the Japanese learner corpus (JLC) project. The students' proficiency level is considered as advanced according to their TOEIC scores (700~750). The learners are supposed to write a 500-word essay with given topics during an obligatory course. In reality, the range is 400~500, thus the mean size of each essay is about 450 words. These data are collected by Professor. Hong of Busan University of Foreign Studies.

2.3 Corpus compilation

The study analyses 4-grams in EFL Japanese learners' written production (JLC) and native English essays (LOCNESS). Due to this, 366 academic essays are used to compile the JLC (Table 1).

Table 1. Size of the corpora

Corpus	Tokens	Number of texts
Japanese learner corpus (JLC)	199239	366
LOCNESS	324008	411

They are collected from several universities in Japan. The design criteria for JLC are the same as those of the ICLE (international corpus of learner of English) project conducted by Granger (Table 2).

Table 2. Design criteria of corpora used in the study

Features	Category	Attributes	
		Japanese learner corpus	LOCNESS
Language	Mode	Written	Written
	Genre	Academic essay	Academic essay
	Style	Argumentative	Argumentative
	Topic	Given topics	Given topics
Learners	Age range	20-30 years	University students
	Proficiency level	Advanced	Native speakers
	Mother tongue	Japanese	English
	Learning context	EFL	--

(Continuation table)

Task	Data collection	Cross-sectional	Cross-sectional
	Task setting	Untimed/timed	Untimed/timed
	Elicitation	Prepared	Prepared
	Technicality	Non-technical	Non-technical

2.4 N-gram lists

The study adopts norminalised lists of n-grams in order to determine the specific number of grams. It is key to set cut-off points indicating threshold frequency for analyzing recurrent patterns of n-grams from different sizes of corpora. In the previous studies, there is no agreement in this aspect (Chen & Baker, 2010). In general, three sorts of cut-off frequency need to be considered: the number of continuous sequences, threshold frequency in an n-gram list, and dispersion of n-grams in a text. For continuous words number, 4-grams are considered to be the ideal number of words according to Cortes (2006) and Adel and Erman (2012). One reason for using 4-grams in this study is that many other researchers have done with 4-grams so that the results can be compared to others. Another reason is that 4-grams contain 3-grams and are more numerous in the list than 5-grams (Cortes, 2004; Hong, 2013). Then threshold frequency should use to cut-off at some point in the full list of 4-grams in order to identify conventional words. Many studies concerning cut-off point with different criteria. For this study, Chen & Baker's (2010) is adopted with 25 times per million words.

The last thing, the dispersion of the threshold frequency (25 per million words) was set at 3 texts in this study. According to Biber and Barbieri (2007), n-grams have to occur in different texts (at least 3~5) for small corpora. Therefore the cut-off point was set at frequency 8, text 3 for LOCNESS, frequency 5, text 3 for JLC (Table 3).

Table 3. Threshold frequency for this study

Corpus	Raw frequency	Frequency	Dispersion
JLC	4.98	5	3
LOCNESS	8.10	8	3

2.5 Grammatical categorization

The study analyzes n-gram lists from a grammatical perspective. Biber etc., (1999) grammatical categorization is used as a basis of this study category, since his categorization is widely used for n-gram analysis so that contrastive analysis can easily be achieved. Table 4 shows the grammatical categories in the study.

Table 4. Grammatical categories in the study

Category	Subcategory	Example
Noun-related category	Noun phase (NP)	The use of the
Preposition-related category	Prepositional phrase + NP (PP)	In the field of
Verb-related categories	Passive (PA)	Is based on the
	be+ NP/AP*/PP (BE)	Is one of the
	Verb+ NP/PP (VN)	Use the credit card/need to use the
	Modal (MO)	Would spend less time
	-ing (ING)	Watching movies at home
	To-infinitive (TO)	To think about the
Clause-related categories	Anticipatory-it (Ant-it)	It is difficult to
	Conjunction (CO)	When I was young
	NP/Pronoun+rerb(NV)	We need to think
Others (OT)		There has been a/not good at English/ not too much to/difficult for us to

3. Results and discussion

3.1 4-gram lists

The analysis of Japanese learners' corpus focuses on 4-grams with the cut-off points (25 per million and at least 3 texts). As Table 5 shows, the Japanese learners use more 4-grams in academic essays than native English learners.

Table 5. Frequencies of 4-gram lists

	JLC	LOCNESS
	With the cut-off point	With the cut-off point
Token	5861	2389
Type	594	178
TTR	10.13	7.45

Since the two corpora have the same design criteria, the difference can be summarized as Japanese learners use more 4-grams. That is, they are likely to depend on more recurrent continuous sequences of words than native speakers. Relative frequencies of the token with cut-off point are 3.96% (JLC) and 0.60% (LOCNESS). The results of type and TTR for these two corpora are similar to token.

3.2 Grammatical categories

The study analyzes Japanese learners' use of 4-grams in terms of grammatical structures. Details are shown in Table 6.

Table 6. Distribution of grammatical categories in each corpus

	Sub-category	JLC			LOCNESS		
		Token	Type	TTR	Token	Type	TTR
Noun-related category	NP	1202	107	0.0890	1007	71	0.0705
		20.51%	18.01%		42.15%	39.89%	
Preposition-related category	PP	703	71	0.1010	743	49	0.0659
		11.99%	11.95%		31.10%	27.52%	
Verb-related category	PA	24	4	0.1667	10	1	0.1000
		0.41%	0.67%		0.42%	0.56%	
	BE	341	37	0.1085	95	7	0.0737
		5.82%	6.23%		3.98%	3.93%	
	VN	693	76	0.1097	54	4	0.0741
		11.82%	12.79%		2.26%	2.25%	
	MO	118	14	0.1186	97	9	0.0928
		2.01%	2.36%		4.06%	5.06%	
	TO	262	20	0.0763	40	4	0.1000
		4.47%	3.37%		1.67%	2.25%	
Clause-related category	IT	325	32	0.0985	101	10	0.0990
		5.55%	5.39%		4.23%	5.62%	
	CO	596	73	0.1225	83	7	0.0843
		10.17%	12.29%		3.47%	3.93%	
	NV	1332	132	0.0991	105	11	0.1048
		22.73%	22.22%		4.40%	6.18%	
Others	OT	265	28	0.1057	54	5	0.0926
		4.52%	4.71%		2.26%	2.81%	
Total		5861	594	0.1013	2389	178	0.0745

A distinctive feature is that NV (noun/pronoun + verb) 22.73% is the most frequently used 4-grams for Japanese learners while NP (noun phrase) 42.15% for native speakers. Moreover, NP (20.51%) is the second most commonly used patterns for Japanese learners while PP (prepositional phrase) (31.10%) for native speakers.

3.2.1 Noun/pronoun + verb

I think it is a government's crime and it should be judged as same as the usual crime.

3.2.2 Noun phrase

It has even been attempted to build more roads for the cars to travel on, but it seems that the number of cars increases on the road increase, making the problem the same as it was before.

3.2.3 Prepositional phrase

Another rule is at the end of the race, come in the pit slowly and do not hit the person in front of you.

This feature indicates an overuse of NV among Japanese learners. They tend to use pronoun I (e.g. I think it is / I would like to) to express their opinions and views. On the other hand, native speakers' use of PP (31.10%) shows a similar frequency

pattern to the academic genre in the native reference corpus : JLC 11.99%, LOC 31.10%, LSWE (Longman Spoken and Written English) 33%.

From the perspective of language variety, the 4-grams of JLC and LOCNESS do not have much difference. However, specific TTRs in each category have a different pattern. First, the TTR of PA in the JLC is much higher LOCNESS. This result may be influenced by its low frequency of tokens and types. Second, the category of TO (to- infinitive) in the JLC is somewhat lower than LOCNESS. Thus, Japanese learners prefer some particular patterns of "to do…" than others while native speakers use several different patterns.

4. Conclusion

The descriptive patterns for Japanese learners and native speakers present how they use 4-grams differently. On the basis of grammatical pattern analysis, Japanese learners has a trend to overuse NV phrases and NP comes the second frequently use while native speakers prefer conventional NP and PP phrases. One of the reasons can owe to their writing habits that to express themselves with "I think…". The TTRs also indicate that native speakers are likely to use particular patterns of NP and PP in their writings, but Japanese learners use various 4-grams that is created by themselves or some free combination.

With this regarding, teachers need to think of a method in which how the formulaic patterns can be taught and acquired by learners appropriately and how to encourage students to use it. According to Schmidt's Noticing Hypothesis, language learners are limited in what they are able to notice and the main determining factor is attention. Teachers should adopt some activities or tasks to rise learners' awareness of the formulaic patterns. When the teacher or one learner use the formulaic patterns first, other learners tend to use it again in their production. Given this, learners can use and acquire the patterns.

There are some limitations of this study. First, the grammatical category is not so clear for Japanese learners 4-grams. A number of confusing grammar patterns are set to OT (others). More detailed subcategories are needed to add to the present one. The other issue is that the proficiency level of learners needs to be clearer since advanced level may not be defined properly.

References

1. Adel A, Erman B. Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. English for Specific Purposes 2012; 31: 81-92.
2. Biber D, Conrad S, Cortes V. If you look at...: Lexical bundles in university teaching and textbooks. Applied Linguistics 2004; 25: 371-405.
3. Biber D, Johansson S, Leech G, *et al*. Language grammar of spoken and written English. London: Longman; 1999.
4. Cheng W, Greaves C, Sinclair, J. M., *et al*. Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. Applied Linguistics 2008; 30(2): 236-252.
5. Cortes V. Lexical bundles in published and student disciplinary writing: Examples from history and biology. English for Specific Purposes 2004; 23: 397-423.
6. Cortes V. Teaching lexical bundles in the disciplines: An example from a writing intensive history class. Linguistics and Education 2006; 17+391-406.
7. Granger S, Meunier F. Phraseology: An interdisciplinary perspective. Amsterdam: John Benjamins; 2008.
8. Hoey M. A world beyond collocation: New perspectives on vocabulary teaching. In M. Lewis (Ed.). Corpora and language learners. Amsterdam: John Benjamins; 2004.
9. Hong S. EFL learners' consciousness-raising through a corpus-based approach. English Teaching 2010; 65(1): 57-86.
10. Hong S. An n-gram analysis of maritime English. The Journal of Linguistic Science 2012; 61(2): 283-328.
11. Pawley A, Syder F. Two puzzles for linguistic theory native like selection and native like fluency. In J. C. Richards & R. W. Schmidt (Eds.). Language and communication. London: Longman; 1983.
12. Schmidt R. The role of consciousness in second language learning. Applied Linguistics 1990; 11: 129-158.
13. Schmidt R. Deconstructing consciousness in search of useful definitions for applied linguistics. AILA Review 1994; 11: 11-26.
14. Scott M. WordSmith tool (version 5.0) [Computer software]. Oxford: Oxford University Press; 2010.
15. Sinclair J. Corpus, concordance, collocation. Oxford: Oxford University Press; 1991.
16. Wray A. Formulaic language: Pushing the boundaries. Oxford: Oxford University Press; 2008.