Trends

JALTCALL Trends
ISSN 2982-3749, 2(1), 103497 (2026)
https://doi.org/10.29140/jct.v2n1.103497

Castledown

A comparative study of key 4-gram lexical bundles in Japanese L2 English and L1 English spoken monologues

Martin Spivey
Akita University, Japan
mart.spiv@gmail.com

Numerous studies have investigated the language production of foreign language learners through corpus methods (Altenberg & Granger, 2001; Bychkovska & Lee, 2017; Granger & Paquot, 2009). However, to date, most empirical investigations have focused on written production with limited corpus-based research into how learners speak in their foreign language. This paper reports on a register analysis of 4-gram lexical bundles produced by Japanese L2 English speakers and L1 English speakers using data taken from the ICNALE Spoken Monologues corpus (Ishikawa, 2023). Key 4-grams are extracted via Sketch Engine (Kilgarriff et al., 2014) and analyzed through Biber et al.'s (2004) structural and functional taxonomies. Results show that there are similarities between the two groups with regards to both structural and functional bundle types. However, L1 English speakers display a higher usage of noun/preposition-based bundles and referential bundles. Furthermore, the Japanese L2 English speakers produce various erroneous bundles. This study highlights key differences in the language production of L1 English and Japanese L2 English speakers and can be informative for English as a Foreign Language (EFL) teaching practitioners.

Keywords: learner corpora, lexical bundles, second language learning, spoken monologues

Introduction

A learner corpus is "...a collection of machine-readable texts consisting in representative samples of the language written and/or spoken by learners of an additional language..." (Meunier, 2020, p. 23). Some studies have utilized learner corpora to investigate formulaic language involving (dis)continuous word sequences of varying lengths (typically 2–6 words) in a corpus, sometimes known as "lexical bundles" (Biber et al., 2004; Chen & Baker, 2016; Hyland, 2008).

This study attempts to gain a greater understanding of the differences between Japanese L2 English speakers and L1 English speakers when speaking English,

JALTCALL
Trends
Vol. 2 No. 1
2025 JALTCALL
Conference Papers
Special Issue

specifically during spoken monologues. Two sub-corpora were compiled from the ICNALE Spoken Monologues corpus (Ishikawa, 2023). A key 4-gram comparative analysis was conducted, and results show largely similar structural and functional compositions with some noticeable differences. L1 English speakers produced a greater proportion of noun/prepositional phrase (NP/PP) bundles while Japanese L2 English speakers showed an over-reliance on discourse-organizing (DO) bundles such as *I have two reasons* and *with the statement because*. The findings inform EFL pedagogy, highlighting the need for Japanese students of L2 English to develop their use of more complex NP/PP bundles to sound more proficient. The next section provides a brief overview of studies focusing on lexical bundles in spoken learner language.

Lexical bundles in spoken learner language

Lee and Zipagan (2018) investigated bundles in Korean EFL student oral proficiency interviews through a comparison with L1 English speakers completing the same assessment. They discovered that, while both groups showed a tendency to use verb-phrase (VP) based bundles and stance expressions (SE), L1 English speakers produced a greater variety of bundles as well as more NP/PP bundles.

In another study, Sánchez Hernández (2013) examined the 4-gram bundle production of Spanish L2 English majors and L1 English speakers through a corpus of texts from various oral English tasks. The data show that the L2 speakers accounted for a larger percentage of VP bundles and a smaller proportion of NP/PP bundles. In addition, the L1 English speakers created the smallest number of both bundle types and tokens.

In a study on vocabulary production, Shirato and Stapleton (2007) compared bundles in a small corpus of conversations by L1 Japanese adult EFL students with a British National Corpus sub-corpus. The researchers discovered that the foreign language learners rarely (if ever) use certain phrases for expressing vagueness (e.g. *something like that*), as well as there being an absence of discourse markers such as *you know what I mean,* highlighting key differences in spoken fluency.

Lastly, Sitler and Spivey (2025) compared 4-gram bundles produced by Japanese L2 English speakers and L1 English speakers in the ICNALE Written Essays and Spoken Monologues corpora through a multi-modal register analysis. Some key findings include the discovery that Japanese students produce a greater variety and higher frequency of bundles when writing in English compared to when speaking. Japanese L2 English speakers also use a larger proportion of VP bundles than any other group, as well as producing significantly fewer prompt-based bundles than L1 English speakers in the monologue task.

The present study

A small-scale study was designed to answer the following research question:

1. What are the key differences between Japanese L2 English speakers and L1 English speakers in the production of 4-gram lexical bundles during spoken monologue tasks?

Methods

Two sub-corpora of the ICNALE Spoken Monologues corpus were used for this investigation: Japan Spoken Monologues (henceforth "L2ENG-SM") and English Native Speaker Spoken Monologues (henceforth "L1ENG-SM"). Spoken data were collected via telephone using a server-based automatic recording tool. 150 Japanese students (CEFR level A2~B2+) from several universities contributed to the L2ENG-SM while there were 150 native English-speaking teachers, university students, and adults in the L1ENG-SM. Participants had 60 seconds each time to produce two argumentative monologues on separate topics; college students having a part-time job and banning smoking in restaurants, and repeated the task for each topic resulting in four files per speaker (Ishikawa, 2023). Table 1 shows the data of the two sub-corpora.

Table 1. The L2ENG-SM and L1ENG-SM sub-corpora

	L2ENG-SM	L1ENG-SM
Tokens	48,469	105,100
Files	600	600
Average tokens per file	80	175

Lexical bundles of four words in length were chosen for analysis as this is typical in several learner corpora studies examining bundle production (Biber et al., 2004; Chen & Baker, 2010; Hyland, 2008). The key lexical bundles were computed through the corpus software Sketch Engine (Kilgarriff et al., 2014). To identify key bundles, the two corpora were compared to determine which bundles feature in one corpus but infrequently (or not at all) in the other. The initial searches returned 2530 key bundles for the L2ENG-SM corpus and 4507 for the L1ENG-SM corpus. Bundles with a keyness score of at least 100 were selected, reducing the bundles to 163 and 62 for the L2ENG-SM and L1ENG-SM corpora, respectively. The bundles were then analyzed for their structures and functions using Biber et al.'s (2004) taxonomies (see Appendices A & B).

Results

Table 2 presents the total number of bundle types and token raw frequencies for both the structures and functions. First, structurally, the data show that in both corpora, most bundles are VP-based. L1ENG-SM has a greater proportion of VP bundle type and token frequencies than L2ENG-SM. The frequency of NP/PP bundles produced by L1 English speakers is just over a quarter of the total, though Japanese speakers show a similar rate (23.2%). The key difference in the structures is the frequency of dependent clauses (DC). Proportionally, DC are roughly half as frequent in L1ENG-SM as they are in L2ENG-SM.

Table 2. Key 4-gram lexical bundles

		Structural types					Functional types			
Corpus		Type Freq.	%	Token Freq.	%		Type Freq.	%	Token Freq.	%
L2ENG-SM	VP	103	63.2	799	67.0	SE	112	68.7	814	68.2
	DC	21	12.9	117	9.8	DO	16	9.8	161	13.5
	NP/PP	39	23.9	277	23.2	RF	35	21.5	218	18.3
	Total	163	100	1193	100	Total	163	100	1193	100
L1ENG-SM	VP	42	67.7	782	68.8	SE	41	66.1	776	68.3
	DC	3	4.9	57	5.1	DO	1	1.6	26	2.3
	NP/PP	17	27.4	297	26.1	RF	20	32.3	334	29.4
	Total	62	100	1136	100	Total	62	100	1136	100

Functionally, both corpora display a high proportion of SE bundles which could be expected in an argumentative spoken task in which speakers need to express their opinions. Intriguingly, L1ENG-SM featured only one discourse organizer (DO) (*on the other hand*), whereas there are multiple uses of DO in L2ENG-SM (e.g. *I have two reasons, with the statement because*). Greater reliance on DO in L2ENG-SM is likely to be because of instruction given to university students advising regular use of sign-posting phrases for coherence. Finally, regarding RF bundles, the table highlights a significant difference in usage levels between the two corpora. Just two RF bundles appear in the top 20 key 4-gram list of L2ENG-SM (*to do part-time job, to work part-time job*) whereas there are six in the L1ENG-SM, including *a part-time job while, for students to have,* and *to have a job*.

The analysis also revealed the number of bundles in the L2ENG-SM that included language errors. 37 bundles (22.7%) were found to have some form of grammatical error with the majority missing a preposition or article, such as *I agree this statement, to do part-time job, is bad for health*. While these errors do not necessarily hinder comprehension, they highlight a lack of proficiency on the part of some learners.

Discussion

Despite a lack of studies which apply the same methodology, some comparisons between this study and similar research can be made. As in Lee and Zipagan (2018), the data show both groups produced a majority of VP and SE bundles, as well as evidence of L1 English speakers producing a greater proportion of NP/PP bundles. Interestingly, the Japanese L2 English speakers have a much greater number of bundle types in the list yet have a lower average token frequency. One explanation for the former is the high number of grammatical errors made by some students, leading to a wider variety of (erroneous) bundles than for L1 English speakers. While Shirato and Stapleton (2007) showed that L1 English speakers produced more vague bundles than L1 Japanese speakers, only one such bundle was found in the former group's key bundles list (*and things like that*), which may be due to the prominence of topic-related bundles in a task with only two topics under discussion. It is also of note that in the present study, L1 English speakers were slightly more reliant on VP bundles in

their key 4-grams than Japanese L2 English speakers, the opposite effect found in the bundle production of both groups in Sitler and Spivey (2025). Regardless, L1 English speakers still produce a higher proportion of NP/PP bundles in both types and tokens, pointing to a greater level of proficiency in their spoken production.

JALTCALL Trends
Vol. 2 No. 1
2025 JALTCALL Conference Papers
Special Issue

Conclusion

This study provides greater insight into how Japanese L2 English speakers compare with L1 English speakers when producing 4-gram lexical bundles in argumentative spoken monologue tasks. A key 4-gram analysis shows that even though both groups display structural and functional similarities, with respect to key bundles, Japanese L2 English speakers produce fewer NP/PP and RF bundles than L1 English speakers. Additionally, just over a fifth of key bundles featured a grammatical error of some kind.

This study is informative yet not without limitations. Firstly, the data analysis was conducted solely by this researcher and joint analysis may have strengthened the veracity of the results. In addition, the corpora are relatively small therefore any generalizations of Japanese EFL students or L1 English speakers cannot be confidently made. Despite these limitations, the study provides some insight into how L1 Japanese speakers compare to L1 English speakers when performing spoken monologues in English and can inform practitioners of English for Academic Purposes in the Japanese university context, providing food for thought when creating curricula and lesson plans designed to develop students' spoken fluency.

References

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics, 22*(2), 173–195. https://doi.org/10.1093/applin/22.2.173

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

Bychkovska, T., & Lee, J. J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes, 30,* 38–52. https://doi.org/10.1016/j.jeap.2017.10.008

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Academic Writing, 14*(2), 30–49. https://doi.org/10.64152/10125/44213

Chen, Y.-H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics, 37*(6), 849–880. https://doi.org/10.1093/applin/amu065

Granger, S., & Paquot, M. (2009). Lexical verbs in academic discourse: A corpus-driven study of learner use. In M. Charles, S. Hunston, & D. Pecorari. (Eds.), *Academic writing: At the interface of corpus and discourse.* Bloomsbury Academic. https://doi.org/10.5040/9781474211703

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001

Ishikawa, S. (2023). The ICNALE guide: An introduction to a learner corpus study on Asian learners' L2 English. Routledge.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Lee, K. R., & Zipagan, M. N. (2018). Korean English learners' use of lexical bundles in speaking. *The Journal of Asia TEFL, 15*(2), 276–291. https://doi.org/10.18823/asiatefl.2018.15.2.2.276

Meunier, F. (2020). Introduction to learner corpus research. In N. Tracy-Ventura, & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 23–36). Routledge. https://doi.org/10.4324/9781351137904

Sánchez Hernández, P. (2013). Lexical bundles in three oral corpora of university students. *Nordic Journal of English Studies, 12*(1), 187–209. https://doi.org/10.35360/njes.281

Shirato, J., & Stapleton, P. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research, 11*(4), 393–412. https://doi.org/10.1177/1362168807080960

Sitler, T., & Spivey, M. (2025). "I agree with this": A multi-modal register analysis of lexical bundles in the ICNALE. *Asia Pacific Journal of Corpus Research, 6*(1), 25–45. https://doi.org/10.22925/apjcr.2025.6.1.25

Appendix A

A condensed summary of Biber et al.'s (2004) Structural Taxonomy of Lexical Bundles

1. Verb phrase (VP) fragments
   (*you don't have to, it's going to be, is based on the etc.*)
2. Dependent clause (DC) fragments
   (*I don't know why, if you want to, that there is a etc.*)
3. Noun phrase and prepositional phrase (NP/PP) fragments
   (*one of the things, a little bit more, as far as the etc.*)

Appendix B

A condensed summary of Biber et al.'s (2004) Functional Taxonomy of Lexical Bundles

1. Stance expressions (SE)
   (*I don't know if, it is important to, is going to be etc.*)
2. Discourse organizers (DO)
   (*if you look at, want to talk about, on the other hand etc.*)

3. Referential expressions (RF)
   (*one of the most, there's a lot of, at the same time etc.*)

JALTCALL
Trends
Vol. 2 No. 1
2025 JALTCALL
Conference Papers
Special Issue

Author biodata

Martin Spivey is Assistant Professor of EAP at Akita University in Tohoku, Japan. He has an MA in TESOL from the University of Birmingham and over 20 years of experience in the foreign language classroom. His research interests include discourse analysis, data-driven learning, and learner corpora.