Asia Pacific Journal of Corpus Research
Vol. 4, No.2, 1-39

http://www.icr.or.kr/apjcr
e-ISSN 2733-8096
https://doi.org/10.22925/apjcr.2023.4.2.1

Formulaic Language Development in Asian Learners of English: A Comparative Study of Phrase-frames in Written and Oral Production

Yoon Namkung
(Georgia State University)
Ute Römer
(Georgia State University)

Namkung, Y., & Römer, U. (2023). Formulaic language development in Asian learners of English: A comparative study of phrase-frames in written and oral production. *Asia Pacific Journal of Corpus Research, 4*(2), 1-39.

Recent research in usage-based Second Language Acquisition has provided new insights into second language (L2) learners' development of formulaic language (Wulff, 2019). The current study examines the use of phrase-frames, which are recurring sequences of words including one or more variable slots (e.g., *it is * that*), in written and oral production data from Asian learners of English across four proficiency levels (beginner, low-intermediate, high-intermediate, advanced) and native English speakers. The variability, predictability, and discourse functions of the most frequent 4-word phrase-frames from the written essay and spoken dialogue sub-corpora of the International Corpus Network of Asian Learners of English (ICNALE) were analyzed and then compared across groups and modes. The results revealed that while learners' phrase-frames in writing became more variable and unpredictable as proficiency increased, no clear developmental patterns were found in speaking, although all groups used more fixed and predictable phrase-frames than the reference group. Further, no developmental trajectories in the functions of the most frequent phrase-frames were found in both modes. Additionally, lower-level learners and the reference group used more variable phrase-frames in speaking, whereas advanced-level learners showed more variability in writing. This study contributes to a better understanding of the development of L2 phraseological competence.

Keywords: Usage-Based SLA, Phraseological Competence, Phrase-Frames, Learner Corpora, English as a Foreign/Second Language

1. Introduction

Corpus-based research on language has demonstrated that multi-word sequences are essential units of language representation. Contrary to the traditional view of language, which sees lexis and grammar as separate components of language, research on phraseology views them together by highlighting their interaction with each other (Römer, 2009). Corpus researchers have tried to better understand second language (L2) learners' use of phraseological units by (1) comparing L2 learners' production of formulaic sequences with those of native speakers (e.g., Nekrasova-Beker, 2009; O'Donnell, Römer, & Ellis, 2013), (2) examining the developmental patterns of L2 learners' use of phraseological items (e.g., Chen & Baker, 2010; Garner, 2016; Nekrasova-Beker, 2021; Tan & Römer, 2022), and (3) conducting contrastive analyses of L2 phraseology of learners from different first language (L1) backgrounds; (e.g., Juknevičienė & Grabowski, 2018; Paquot, 2013).

In line with previous L2 phraseology research, the current study explores phraseological units produced by L2 learners of English by adopting a phrase-frame (hereafter p-frame) approach. A p-

frame is a semi-fixed multi-word sequence that includes one or more variable slots marked by "*" which are filled with so-called "variants" (e.g., *it is * that*, frequent variants: *clear*, *obvious, true*). Thus, p-frames are more flexible in their usage than fixed sequences of words, such as n-grams or lexical bundles (e.g., *it is clear that*), which do not allow for internal variation. Despite its described pedagogical significance (e.g., Juknevičienė & Grabowski, 2018; Liu, Jiang, & Du, 2023; Lu, Yoon, & Kisselev, 2018), this particular type of phraseological item has received the least amount of attention in the study of formulaic sequences within learner language (Tan & Römer, 2022). While recent empirical studies have started to investigate L2 learners' use of p-frames, the focus in these studies is predominantly on written rather than oral production. Despite the different mechanisms underlying writing and speaking, existing studies on L2 learner language have not yet examined how these two modes compare with respect to the use of p-frames. Inspired by these research gaps, the present study examines L2 learners' developmental patterns in their use of p-frames and the similarities and differences in their use of p-frames between written and oral production data. Specifically, we investigate to what extent the variability, predictability, and discourse functions of p-frames are different across L2 proficiency levels among L2 learners of English in both written and oral production. Further, we examine how L2 learners use p-frames in written and oral production differently at various proficiency levels.

2. Literature Review

2.1. The Acquisition of Phrase-Frames in Usage-Based Second Language Acquisition

Usage-based language acquisition emphasizes the importance of learning constructions, which are defined as "conventional, learned form-function pairings at varying levels of complexity and abstraction" (Goldberg, 2013, p. 3). Ranging from morphemes (e.g., the suffix –ly in *gladly*) to complex syntactic frames (e.g., the transitive resultative construction), constructions are formulaic or phraseological items that are essential units of language representation. In the field of usage-based Second Language Acquisition (SLA), corpus research contributes to understanding L2 learners' acquisition of constructions. Previous corpus-based research in this area has examined L2 learners' use of formulaic patterns, such as lexical bundles or n-grams (Biber & Barbieri, 2007; Chen & Baker, 2016; Paquot, 2013), collocations (Nesselhauf, 2005), and phrasal verbs (Gilquin, 2015). In addition to these fixed and continuous sequences, however, exploring frequently occurring discontinuous multi-word sequences with item-internal variation could help better understand how "fixed" a sequence is by providing a systematic grouping of related n-grams. Also, in terms of pedagogical aspects, teaching multi-word sequences with internal variations can allow teachers to "introduce more language while lessening the cognitive demand on memory" (Lu, et al., 2018, p. 78). However, compared to the continuous set expressions, these discontinuous multi-word units have to date received less attention in usage-based SLA research.

Recently, empirical studies have begun to investigate L2 learners' use of p-frames (Garner, 2016; Juknevičienė & Grabowski, 2018; Larsson, Reppen, & Dixon, 2022; Nekrasova-Beker, 2021; O'Donnell et al., 2013; Römer & Banerjee, 2017; Tan & Römer, 2022; Xia, Sulzer, & Pae, 2023). Research on the use of p-frames in written production mostly falls under the following topics: (1) the developmental patterns in the use of p-frames in L2 writing, (2) the comparison of p-frames between novice writers (or learners) and expert writers, (3) the use of p-frames in writing by learners of different L1 backgrounds, and (4) the creation of lists that are pedagogically useful for specific written genres. Firstly, research on learners' use of p-frames in writing has highlighted developmental patterns in their usage of p-frames. For instance, using the German subsection of the EF-Cambridge Open Language Database (EFCAMDAT), Garner (2016) found that L1 German learners of English use more

variable, less predictable, and more functionally complex 4-word p-frames as their L2 proficiency increases, as more advanced learners have likely encountered more input of the p-frames with more word types in the variable slot. More recently, Tan and Römer (2022) examined the developmental patterns of 3- and 4-word p-frames produced by Mandarin Chinese learners of English across different proficiency levels. Using the Chinese subsection of the EFCAMDAT and the Corpus of Contemporary American English (COCA) as a reference, the study investigated the variability, predictability, and functions of high-frequency p-frames. The findings revealed that as learner proficiency increases, learners use more variable and less predictable p-frames in their writings. Also, the 4-word p-frames showed more functional variability than the 3-word p-frames, and lower-level learners tended to use more referential expressions than higher-level learners.

Secondly, studies have also examined how the p-frames produced by novice learners or L2 learners differ from those produced by expert writers or L1 speakers. Xia et al. (2023) investigated p-frame use in business emails by business English learners and working professionals. Using learners' business emails written for assignments in EFCAMDAT and a corpus of business emails from the University of California Berkeley Enron Email Analysis Project, the study revealed that the English learners and business professionals used p-frames differently. Specifically, the business professionals used p-frames with a higher degree of variability and adhered more closely to the written conventions of politeness compared to the learner group. In a similar vein, Larsson et al. (2022) analyzed how novice writers (including L1 and L2 speakers of English) and expert writers use p-frames differently in their academic texts for highlighting purposes. After identifying five target p-frames used for highlighting purposes, the study examined and compared the variants that the two groups used in the selected discontinuous sequences. The results showed that the experts generally used more variable fillers in the slots of p-frames than the novice writers, aligning with previous research.

Thirdly, focusing on potential effects of L1 backgrounds, Juknevičienė and Grabowski (2018) compared the structural features of 4-word p-frames in the written texts of Lithuanian and Polish learners of English. Using two sub-corpora in the International Corpus of Learner English (ICLE) and the Louvain Corpus of Native English Essays (LOCNESS) as a reference corpus, the study highlighted similarities in the Lithuanian and Polish learner groups' use of p-frames. Specifically, the shared p-frames were stance or text-organizing devices, which were mostly preferred by the less proficient learners. Also, the study showed that there were L1 transfer effects in both learner groups' use of p-frames. For example, both Lithuanian and Polish learners underused *of*-frames (e.g., *the * of the*) compared to the reference group because prepositions occupy a different place in both languages compared to English. In a pedagogically motivated study, Lu et al. (2018) compiled a list of academic p-frames for research article (RA) introductions. Using a corpus of published RA introduction sections from six social science disciplines, the most frequently occurring 5- and 6-word p-frames were extracted and then rated for their pedagogical value by instructors and student writers. The resulting list of p-frames could serve as a useful source for helping students with their academic writing. Liu et al. (2023) also investigated the structures and functions of commonly used 3-word p-frames in the frequent moves of figure legends (i.e., descriptive statements accompanying a figure) in scientific RAs. Aligning with Lu et al. (2018), the study highlighted the pedagogical value of connecting rhetorical moves and p-frames.

Likely due to the challenges related to compiling and examining spoken corpora, empirical studies on the use of learners' p-frames in oral production are scarce. With the purpose of providing validity evidence for an L2 speaking test, Römer and Banerjee (2017) conducted a phraseological analysis on test-takers' oral responses in the Michigan English Test (MET) and examined how their phraseological competence differs across learner proficiency bands based on the Common European Framework of Reference for Languages (CEFR) scale. Specifically, the study examined the test-takers' use of 3-, 4-, and 5-word n-grams and p-frames in their oral responses. The findings showed that the test-takers' phraseological competence increased as proficiency level increased, with low-level learners producing

more p-frames with hesitation markers (e.g., *erm*) and repetitions. Nekrasova-Beker (2021) also investigated L2 learners' use of p-frames across three proficiency sub-levels (i.e., low-intermediate, mid-intermediate, and high-intermediate). The study examined differences in the variability and functional characteristics of the p-frames in L2 learners' dyadic oral interactions. Within the intermediate level, the results revealed that the patterns utilized by high-intermediate learners were more variable compared to the patterns produced by mid- and low-intermediate learners. Learners also expanded their uses of p-frames from stance expressions to more diverse discourse functions, such as referential expressions in dyadic interactions, as their proficiency level increased.

While these studies have provided valuable insights, we argue that there is a need for additional research on the development of p-frames in L2 learner speech, especially research which includes target language reference data for comparison with the learner production data. This reference data would ideally be collected in the same context or contexts as the learner data so that context and prompt effects on the language produced are reduced to a minimum.

2.2. Differences between L2 Written and Oral Production

Although writing and speaking utilize the same linguistic resources, the two modes are different in terms of how they are perceived and produced (Chan, Verspoor, & Vahtrick, 2015). Specifically, writing allows for planning and editing, while speaking is more spontaneous and usually does not allow the speaker to plan or edit their utterances (Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006). Thus, it can be assumed that multi-word sequences look very different in written and spoken texts. Using a corpus-driven approach, Biber (2009) found that the multi-word sequences typical in conversations are different from those typical in academic writing; patterns in conversations were found to be more fixed, whereas patterns in academic writing consist of invariable function words with an inner slot that can be filled by content words. Regarding the sequences' structure, the study also discovered that the spoken register contains mostly verb-based sequences (e.g., *I don't know why, I thought that was*). Using the same two corpora used in Biber (2009), Gray and Biber (2013) further examined both continuous and discontinuous multi-word sequences in conversations and academic writing. The study highlighted the importance of looking at discontinuous frames (i.e., p-frames) as recurrent continuous frames (here lexical bundles) do not always capture all the potentially relevant recurring sequences in a corpus. Gray and Biber (2013) confirmed that the spoken register relies more heavily on fixed sequences than the written register. The sequences in conversations usually incorporate high-frequency verbs, whereas the sequences in academic writings are mostly composed of function words and the verb *be*. Overall, these studies have contributed to our understanding of how multi-word sequences differ in speaking and writing. More recently, Hwang, Jung, and Kim (2020) examined the differences between young EFL learners' written and spoken production in terms of syntactic complexity. Using a corpus of written and spoken data, the study found that child L2 learners utilized longer sentences, more subordination, more verb phrases, and less coordination in writing than in speaking.

When examining L2 learners' use of formulaic expressions, learner proficiency is a crucial component. For example, focusing on the structures and discourse functions of lexical bundles produced by L2 learners, Chen and Baker (2016) examined argumentative and expository texts written by L1 Chinese learners of English across different proficiency levels using the Longman Learner Corpus. The findings showed that lower proficiency learners tend to show more colloquial and informal features in their writing, such as verb-based sequences. As proficiency level increased, however, the lexical bundles were characterized by a more formal style of academic writing. To our knowledge, there is a lack of studies exploring L2 learners' use of discontinuous sequences in their written and oral productions. P-frames can offer a different perspective from the lexical bundle approach when understanding L2 learners' phraseological competence. Thus, how L2 learners

produce p-frames differently in written and oral productions warrants more investigation.

2.3. The Current Study

It is evident from the literature that there is a need for more research on L2 learners' development of phraseological patterns, especially p-frames, in oral production, as most studies on the topic have focused on writing. Also, as learners show different features in their written and oral production at different proficiency levels, such as lower-level learners producing more spoken-like features in their writing (Chen & Baker, 2016), it is worth examining how L2 learners' p-frame use differs in the two modes. To address these research gaps, the present study adopts a p-frame approach to examine written and oral production data from Asian learners of English across different proficiency levels using similar topics or prompts. The study examines similarities and differences in the variability, predictability, and discourse functions of the most frequently used 4-word p-frames in the speech and writing of L2 learners and L1 reference speakers. The following research questions guided the present study:

1) To what extent do differences exist in the variability, predictability, and discourse functions of p-frames produced by L2 learners across different proficiency levels in written production data?
2) To what extent do differences exist in the variability, predictability, and discourse functions of p-frames produced by L2 learners across different proficiency levels in oral production data?
3) How are L2 learners' use of p-frames different in written and oral production data across different proficiency levels?

3. Methods

3.1. Description of Corpora

Two sub-corpora of the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2023) were used in this study. The data for ICNALE was collected from college and graduate students in 10 Asian countries (i.e., China, Hong Kong, Indonesia, Japan, Korea, Pakistan, the Philippines, Singapore/Malaysia, Taiwan, and Thailand) who spoke English as a foreign or second language, and from native speakers of predominantly British and American English for reference purposes (Ishikawa, 2013).[1] ICNALE consists of four sub-corpora: spoken monologues, spoken dialogues, written essays, and edited essays. The current study used the written essays (WE) corpus of ICNALE to capture L2 learners' written production data and the spoken dialogues (SD) corpus of ICNALE to capture L2 learners' oral production. We chose the spoken dialogues instead of the monologues, as dialogues in interview formats tend to provide more natural and interactive contexts that resemble natural communication.

The WE sub-corpus (v2.4, 2019) consists of 5,600 short essays (200-300 words) about two common topics (Ishikawa, 2013). The first topic asked whether the participants agreed or disagreed with the statement "It is important for college students to have a part-time job;" the second topic asked whether the participants agreed or disagreed with the statement "Smoking should be completely banned at all the restaurants in the country." The participants were given 20 to 40 minutes to write each essay. The SD sub-corpus (v1.2, 2021) consists of 4,250 transcripts of 30-40 minute oral interviews that include (1) a conversation about the participants' English learning experience, (2) two picture descriptions with related questions, (3) two role-plays with related questions, (4) L2 reflections, and (5) L1 reflections

[1] The Written Essays sub-corpus of ICNALE included data from Singaporean learners, but the Spoken Dialogues sub-corpus replaced data from Singaporean learners with data from Malay learners.

(Ishikawa, 2019). The two picture description tasks were related to the two topics in the WE sub-corpus. The participants were asked to describe six pictures of a boy who has a part-time job at a computer shop and a different set of six pictures about a mother with her son, who tells a nearby smoker to stop smoking in the park. The two role-play tasks are also related to the same two topics. Participants were asked to play the role of a college student who needs to persuade their supervisor that students should have part-time jobs, and the role of a customer who needs to persuade a restaurant owner to give them a refund due to too much smoking inside the restaurant. The learners' reflections in their respective L1s and the interviewers' utterances were excluded from our analysis.

The ICNALE data were divided into four proficiency level groups (beginner, low-intermediate, high-intermediate, advanced) and one native English speaker (NES) group. As our study aims to cross-sectionally investigate L2 learners' formulaic language development, ICNALE data from all four learner proficiency levels were used. Table 1 provides an overview of the number of participants, texts, and words in each sub-corpus used in this study. The WE and SE sub-corpora both contained the highest number of texts from high-intermediate learners, with the smallest number of texts coming from NESs. Also, the WE sub-corpus was larger than the SE sub-corpus in terms of the number of participants, texts, and words. Table 2 shows how the proficiency levels of texts in ICNALE correspond to students' iBT Test of English as a Foreign Language (TOEFL) scores.

Table 1. Description of the Written Essays (WE) and Spoken Dialogue (SD) Sub-corpora in ICNALE

Group	Sub-corpus	Number of Participants	Number of Texts	Number of Words
Beginner	WE	480	960	210,822
	SD	66	660	93,205
Low-intermediate	WE	952	1,904	429,836
	SD	89	890	157,640
High-intermediate	WE	936	1,872	439,326
	SD	173	1730	318,100
Advanced	WE	232	464	111,290
	SD	77	770	164,179
Native English Speakers	WE	200	400	88,999
	SD	20	200	45,301
Total	WE	2,800	5,600	1,280,273
	SD	425	4,250	778,425

Table 2. Description of each ICNALE Group's English Proficiency Level (Ishikawa, 2023)

Group	iBT TOEFL Scores
Beginner (A2)	score < 57
Low-intermediate (B1_1)	57 $\leq$ score < 72
High-intermediate (B1_2)	72 $\leq$ score < 87
Advanced (B2)	87 $\leq$ score
Native English Speakers (NES)	-

3.2. Identification and Analysis of P-frames

We used the concordance tool *AntConc* (Anthony, 2022) to automatically extract 4-word p-frames from the ICNALE sub-corpora. Following existing studies on p-frames (Garner, 2016; Nekrasova-Beker, 2021; Tan & Römer, 2022), we only examined 4-word frames that had one inner open slot (e.g., *it is * that*) and selected the 100 most frequent p-frames identified in each of the ten level-specific WE and SD sub-corpora for further analysis (see the Appendix for a complete list of p-frames for each of the five groups in both written and oral production data).

In the WE datasets, we adjusted the token definitions to include apostrophes and hyphens, ensuring that words such as *don't* (with an apostrophe) and *part-time* (with a hyphen) were treated as single

words rather than being separated into two. Thus, the p-frames in our results lists do not contain incomplete words (e.g., *t*) or words that were part of hyphenated compounds (e.g., *part* in *part-time*). If a p-frame overlapped with a phrase from the task prompt and the most frequently used word in the inner slot was the same as in the prompt (e.g., *a part * job, have * part time, banned at * the, restaurants * the country*), the p-frame was removed, as the learner likely borrowed it from the prompt and it may not be evidence of their productive linguistic ability and hence distort the results (see also Paquot, 2013, 2014). Also, only p-frames that occurred in essays written for both prompts were included to minimize prompt effects (e.g., *part * job is* and *I * smoking should* were eliminated). Further, if a p-frame overlapped with another p-frame (e.g., *a * of money* and *a lot * money*), only the more frequent one was retained. Occasional typos in p-frame variants observed in the WE sub-corpus (e.g., "disterbence" instead of "disturbance" in the low-intermediate learners' written production data) were not corrected.

In the SD datasets, hyphens were not included in the token definition, as hyphens were used for undecipherable utterances in the transcriptions (e.g., ----). Apostrophes, however, were included to capture contractions, such as *don't,* as one word. Only p-frames that occurred in both prompts were included, and p-frames that were included only in the introductions or reflections (e.g., *I * speak English, to * in English*) were excluded to minimize the effect of the task prompt. Due to the interactive nature of dialogues, there were high numbers of repetitions (e.g., *we we, I I*) and hesitations (e.g., *uh, mmm,* and *um*). Although analyzing p-frames that include repetition and hesitation markers can provide valuable insights into learners' oral production (Römer & Banerjee, 2017), we followed Nekrasova-Beker (2021) in excluding p-frames with repetition and hesitation markers (e.g., *uh I * I, I I * to*) in the current study, as they were not considered meaningful units. However, p-frames that included such phenomena in the variable slot were retained (e.g., *I think uh it's* included in the p-frame *I think * it's*). Lastly, if a p-frame overlapped with another p-frame, only the more frequent one was retained.

To answer our three research questions, the variability, predictability, and discourse functions of the 100 most frequent p-frames in the learners' written and oral production data were examined. Variability was operationalized as the ratio of variants to p-frames (variant/p-frame ratio, VPR; Römer, 2010), which is comparable to the type-token ratio method used in other p-frame studies (e.g., Gray & Biber, 2013). VPRs are calculated by dividing the number of variant types or slot-fillers by the number of tokens of the p-frame. VPR values range from 0 to 1, with a VPR close to 0 indicating that the p-frame is fixed, and a VPR close to 1 indicating that the p-frame is variable. For instance, in the advanced learners' written production data, the p-frame *agree with * statement* has a low VPR of 0.09 (top variants: *the, this, my*), whereas the p-frame *of the * of* has a high VPR of 0.89 (top variants: *health, disadvantages, benefits, ability, harm, taste, habit*). The current study adopted the five-category thresholds used in Tan and Römer (2022) for the analysis of variability, which is presented in Table 3.

Table 3. Variability Threshold Categories

VPR	Variability
$x \leq 0.20$	Highly Fixed
$0.20 < x \leq 0.40$	Fixed
$0.40 < x \leq 0.60$	Somewhat Variable
$0.60 < x \leq 0.80$	Variable
$0.80 < x$	Highly Variable

To measure the predictability of each p-frame, we used the normalized entropy values (Gries & Ellis, 2015) provided by *AntConc*. Normalized entropy (H_{norm}) is a measure of uncertainty of a probability distribution (Kumar, Kumar, & Kapur, 1986), in our case the distribution of variants in the "*" slot of a p-frame. An H_{norm} value closer to 0 indicates that the variants are unevenly distributed and predictable,

whereas a normalized entropy closer to 1 demonstrates that variants within the slot are evenly distributed and unpredictable. For example, in the beginner learners' oral production data, the p-frame *I * to be* is fairly predictable with an entropy value of 0.35 (top variants: *want, like*), while the entropy value of the p-frame *the * is very* is 0.98 (top variants: *sea, taste, park, woman, sunset*), making this a less predictable p-frame. The current study followed Tan and Römer (2022) and compared the entropy values of a selection of p-frames across proficiency levels and the reference data. As entropy values are p-frame-specific, it would not be meaningful to calculate mean values for a group of p-frames. Also, different from VPR values, which tend to be systematically related to text type and learner proficiency, a high H_{norm} value does not necessarily provide an indication of a learner's proficiency. Therefore, we aimed to see if learners tended to move towards the entropy values of p-frames produced by the NES group.

Finally, a concordance analysis was conducted to examine the primary discourse function of each p-frame. The 100 identified p-frames for each group were classified into four function categories based on the classification system proposed by Biber, Conrad, and Cortes (2004). It is possible for p-frames with semantically unrelated variants to differ in their discourse functions. We categorized each p-frame into the discourse function that explained the majority of its variants, following previous research (Garner, 2016; Tan & Römer, 2022). Table 4 lists the four discourse functions we used, together with examples from our corpus data.

Table 4. Four Primary Discourse Functions Identified in the Current Study

Discourse Function	Description	Examples from the Current Study
Referential Expressions	p-frames referring to physical or abstract entities and identifying their specific attributes	*the * effects of,* *have a * of,* *is one * the*
Stance Expressions	p-frames used to express attitude or evaluation	*think that * is,* *I * with the,* *that * is important*
Discourse Organizers	p-frames that express relationships between parts of the discourse	*at the * time,* *there are * reasons,* *on the * hand*
Special Conversational Expressions	p-frames typically used to directly address the listener/reader	*I * you to,* *thank * very much,* *you are * to*

3.3. Statistical Analysis

Statistical analyses for this study were carried out in R. Following previous research (Garner, 2016; Nekrasova-Beker, 2021; Tan & Römer, 2022), a Kruskal-Wallis H test was conducted to analyze differences in the variability of p-frames across groups because the data were not normally distributed. Post-hoc analyses were also conducted to determine pairwise differences. Furthermore, aligning with previous empirical studies on learners' production of p-frames (Garner, 2016; Nekrasova-Beker, 2021; Tan & Römer, 2022), we used a Pearson's chi-square test to determine whether the distribution of p-frames across the four discourse function categories was significantly different across groups. We also conducted a Wilcoxon Signed-Rank Test to examine differences in variability values between written and oral production data across groups. Effect sizes were calculated[2] with the Alpha level set to .05. Although using only statistical analyses may not fully capture the subtleties of qualitative distinctions in the variability, predictability, and discourse functions of p-frames across different proficiency levels, it allows us to determine differences that are significant at the group level.

[2] The effect sizes for non-parametric tests could be small (0.01<0.06), moderate (0.06<0.14), or large (≥0.14) (Kassambara, n.d.).

4. Results

Tables 5 and 6 provide descriptive statistics for the 100 most frequently used 4-p-frames across learner and NES groups in written and oral production data. For the written data, normalized frequencies are the highest in the beginner group and the NES group. The high-intermediate group demonstrates the lowest mean frequency. However, in oral production, there is a decline in frequency as proficiency level increases, with the advanced learner group being an exception to this trend.

Table 5. Frequencies of the Top-100 4-p-frames across Groups in Written Production (per 100,000 Words)

	Beginner	Low-intermediate	High-intermediate	Advanced	Native English Speakers
M (SD)	18.41 (9.81)	16.50 (8.32)	14.10 (9.40)	15.86 (9.13)	18.09 (9.16)
Median	14.58	13.74	11.36	12.10	14.56
Minimum Frequency	10.07	9.84	7.68	8.92	10.92
Maximum Frequency	59.72	53.43	78.21	68.14	53.87

Table 6. Frequencies of the Top-100 4-p-frames across Groups in Oral Production (per 100,000 Words)

	Beginner	Low-intermediate	High-intermediate	Advanced	Native English Speakers
M (SD)	29.17 (21.57)	25.06 (17.07)	21.96 (14.78)	24.12 (16.41)	17.48 (9.61)
Median	24.37	20.95	16.62	20.07	13.63
Minimum Frequency	15.77	13.47	12.38	14.25	9.74
Maximum Frequency	189.22	127.91	110.78	132.74	75.94

4.1. Variability, Predictability, and Discourse Functions of P-frames in L2 Written Production

To answer our first research question, the variability, predictability, and discourse functions of the 100 most frequent p-frames in each group's written production data were examined. First, regarding the variability of p-frames in the written data, Table 7 shows the distribution of p-frames in each threshold category for each group. There were statistical differences in the variability of p-frames across groups in the written production and the effect size was moderate ($H(4) = 48.15$, $p = .000$, $\eta^2 = .09$). Post-hoc analyses revealed that there were pairwise differences between the beginner and advanced groups ($p = .000$), beginner and NES groups ($p = .013$), low-intermediate and high-intermediate groups ($p = .009$), low-intermediate and advanced groups ($p = .000$), and low-intermediate and NES groups ($p = .000$), and high-intermediate and advanced groups ($p = .016$). The results demonstrated that the variability of p-frames tends to increase as proficiency level increases. The advanced learner group used p-frames with more diverse sets of variants than the NES group, but this result was not statistically significant.

Table 7. Distribution of P-frames s Variability Threshold Categories for Written Production

	Beginner	Low-intermediate	High-intermediate	Advanced	Native English Speakers
Highly Fixed ($x \leq 0.20$)	47	58	43	23	29
Fixed ($0.20 < x \leq 0.40$)	26	19	20	23	34
Somewhat Variable ($0.40 < x \leq 0.60$)	15	17	20	26	17
Variable ($0.60 < x \leq 0.80$)	7	6	16	16	12
Highly Variable ($0.80 < x$)	5	0	1	12	8

To identify potential differences in the predictability of p-frames between groups, we compared normalized entropy (H_{norm}) values across datasets. Table 8 lists the H_{norm} values of the 27 p-frames that occurred in the top-100 lists of all datasets. The NES data is used as a point of comparison. Some of the p-frames in Table 8 are highly predictable with low H_{norm} values in all datasets, for example *I * it is, that * is important,* and *should be * in.* Other p-frames with H_{norm} values closer to 1 are less predictable, including *it is * that, to * in the,* and *and the * of.* With a few exceptions (e.g., *it is * good*), the H_{norm} values are generally lower in the beginner, low-intermediate, and high-intermediate level learners' datasets than the H_{norm} values in the advanced learners' dataset, which are closer to the L1 reference group's H_{norm} values. This indicates that advanced learners produce p-frames that are similar to those of NESs in terms of predictability.

Table 8. Normalized Entropy Values of the P-frames That Appear in All Datasets of Written Production

P-frame	Beginner	Low-intermediate	High-intermediate	Advanced	Native English Speakers
*I think * is*	0.41	0.35	0.37	0.37	0.63
*it is * to*	0.83	0.79	0.84	0.88	0.93
*I * it is*	0.26	0.21	0.24	0.38	0.32
*that * should be*	0.18	0.15	0.21	0.25	0.58
*think that * is*	0.48	0.38	0.49	0.65	0.51
*I * that it*	0.42	0.47	0.65	0.87	0.63
*that * is important*	0.17	0.10	0.12	0.22	0.35
*is very * for*	0.76	0.69	0.71	0.81	0.60
*it is * that*	0.90	0.86	0.88	0.92	0.96
*to * in the*	0.83	0.81	0.86	0.88	0.97
*agree with * statement*	0.64	0.71	0.59	0.66	0.74
*if * want to*	0.73	0.66	0.65	0.88	0.81
*it is * good*	0.70	0.60	0.53	0.90	0.48
*is not * to*	0.89	0.83	0.92	0.97	0.90
*in the * and*	0.92	0.91	0.90	0.97	0.92
*is very * to*	0.89	0.83	0.85	0.95	0.93
*in the * of*	0.96	0.96	0.93	0.97	0.91
*with the * that*	0.75	0.74	0.72	0.70	0.80
*it is * important*	0.77	0.66	0.68	0.83	0.78
*a good * for*	0.90	0.88	0.90	0.88	0.93
*for the * of*	0.93	0.82	0.85	0.89	0.91
*will be * to*	0.64	0.74	0.75	0.78	0.91
*do not * to*	0.74	0.68	0.80	0.81	0.76
*the * and the*	0.98	0.96	0.93	0.97	0.99
*should be * in*	0.54	0.40	0.34	0.53	0.31
*it is * a*	0.59	0.69	0.72	0.86	0.94
*and the * of*	1.00	0.96	0.96	0.97	0.96

Table 9 shows the results of the discourse function analysis and shows how the 100 most frequent

p-frames in each dataset are distributed across the four discourse functions. Stance expressions and referential expressions were most frequent in all datasets. There was only one p-frame in the beginner group that was classified as a special conversational expression (i.e., *you * do it*). The beginner group used more stance expressions than referential expressions, while the low-intermediate, high-intermediate, and advanced learner groups, and the L1 reference group used more referential expressions than stance expressions. However, no statistical differences in the distribution of discourse functions were found by group ($\chi^2(12) = 19.22$, $p = .083$).

Table 9. Distribution of P-frames across the Four Discourse Functions in Written Production

	Beginner	Low-intermediate	High-intermediate	Advanced	Native English Speakers
Stance Expressions	53	47	33	40	47
Referential Expressions	41	49	63	57	53
Discourse Organizers	5	4	4	3	0
Special Conversational Expressions	1	0	0	0	0

4.2. Variability, Predictability, and Discourse Functions of P-frames in L2 Oral Production

To answer our second research question, the variability, predictability, and discourse functions of the 100 most p-frames in oral productions were examined. Table 10 provides the number of p-frames in each threshold category for each group. The differences in the variability of p-frames across groups in the oral production was statistically significant and the effect size was moderate ($H(4) = 51.13$, $p = .000$, $\eta^2 = .10$). Specifically, there were significant differences between the beginner and NES groups ($p = .000$), the low-intermediate and NES groups (p = .000), the high-intermediate and NES groups ($p = .016$), and the advanced and NES groups ($p =. 000$). Contrary to the written data, the results showed that variability did not increase as learner proficiency increased. However, all the learner groups used significantly more fixed p-frames than the NES group.

Table 10. Distribution of P-frames across Variability Threshold Categories for Oral Production

	Beginner	Low-intermediate	High-intermediate	Advanced	Native English Speakers
Highly Fixed ($x \leq 0.20$)	28	28	33	33	9
Fixed ($0.20 < x \leq 0.40$)	28	32	37	34	26
Somewhat Variable ($0.40 < x \leq 0.60$)	20	20	15	17	22
Variable ($0.60 < x \leq 0.80$)	15	12	15	9	19
Highly Variable ($0.80 < x$)	9	8	0	7	24

As for the predictability of p-frames, Table 11 displays the H_{norm} values for the 24 p-frames that occurred in the top-100 lists of all datasets. As in the analysis of the written data, the NES data is used as a point of comparison. In Table 11, we see that, different from what was observed in the written data, H_{norm} values do not tend to become closer to the NES group for most of the p-frames (including *in the * and, I * want to, I would * to,* and *have a * of*). The learner groups' p-frames displayed relatively similar H_{norm} values which did not increase with proficiency.

Table 11. Normalized Entropy Values of P-frames That are Shared Across Datasets in Oral Production

P-frame	Beginner	Low-intermediate	High-intermediate	Advanced	Native English Speakers
*a few * ago*	0.27	0.16	0.15	0.23	0.17
*I think * is*	0.75	0.73	0.67	0.68	0.95
*so I * to*	0.74	0.73	0.74	0.73	0.97
*I * to go*	0.73	0.74	0.81	0.79	0.92
*in the * and*	0.88	0.87	0.89	0.87	0.99
*I * like to*	0.64	0.52	0.43	0.62	0.67
*a * of people*	0.24	0.33	0.15	0.23	0.54
*the * of the*	0.99	0.95	0.94	0.90	1.00
*and I * to*	0.87	0.83	0.79	0.83	0.98
*I * want to*	0.58	0.68	0.63	0.68	0.93
*to the * and*	0.94	0.88	0.82	0.84	0.94
*when I * a*	0.52	0.41	0.40	0.44	0.59
*the * and he*	0.98	0.95	0.96	0.93	0.97
*to * in the*	0.89	0.86	0.78	0.82	1.00
*he * to the*	0.88	0.84	0.71	0.71	0.77
*I would * to*	0.37	0.33	0.32	0.37	0.73
*my * and I*	0.93	0.92	0.78	0.81	0.67
*have a * of*	0.31	0.44	0.27	0.23	0.77
*to go * the*	0.00	0.29	0.33	0.25	0.62
*the * and the*	0.99	0.96	0.96	0.97	0.99
*I think * should*	0.89	0.92	0.77	0.82	0.90
*and he * to*	0.97	0.90	0.89	0.91	0.92
*the * and I*	0.99	0.97	0.97	0.96	0.98
*there * a lot*	0.78	0.57	0.71	0.70	0.99

Table 12 shows the distribution of p-frames across the four discourse functions by group in oral production data. Stance expressions were most frequently used by all learner groups, while the NES group used more referential expressions than stance expressions in their oral production data. Discourse organizers and special conversational expressions were rare in the data from all groups. Overall, there were no significant group differences in the distribution of discourse functions ($\chi^2(12) = 12.46$, $p = .410$).

Table 12. Distribution of P-frames across the Four Discourse Functions in Oral Production

	Beginner	Low-intermediate	High-intermediate	Advanced	Native English Speakers
Stance Expressions	57	61	60	56	44
Referential Expressions	42	37	39	43	52
Discourse Organizers	1	1	0	0	1
Special Conversational Expressions	0	1	1	1	3

4.3. Comparison between L2 Written and Oral Production Data

To answer the third research question, which investigated potential differences in how learners use p-frames differently in their written and oral production, we conducted a combination of quantitative and qualitative analyses. We first determined the frequencies of overlapping p-frames in

each group's written and oral data and compared them across levels. The results in Table 13 indicate that generally, as proficiency level gets higher, the number of overlapping p-frames between written and oral production goes down. To be specific, for the beginner and low-intermediate groups, there were 22 and 23 overlapping p-frames between written and oral datasets, whereas the advanced learners and the L1 reference group had 13 and 15 overlapping p-frames, respectively. In other words, lower-level learners showed fewer register differences between the modes of writing and speaking than the higher-level learners and L1 speakers.

Table 13. Numbers of Overlapping P-frames in Written and Oral Production

Group	Number of Overlapping P-frames in Written and Oral Production	Overlapping P-frames
Beginner	22	*I think * is, it is * to, I * it is, the * of the, I think * should, to * in the, I * with this, if * want to, it is * good, a * of people, a lot of people, in the * and, want to * a, is a * of, the * is not, have a * of, in the * is, agree with * opinion, I * agree with, the * and the, is very * and, there are * people*
Low-intermediate	23	*the * of the, it is * to, I think * is, I * it is, is very * for, is not * for, I think * should, to * in the, have a * of, it is * good, is * good for, is a * of, have * lot of, a * of people, if * want to, in the * and, want to * a, the * and the, are a * of, the * is not, there * a lot, are * lot of, there are * lot*
High-intermediate	18	*the * of the, it is * to, I think * is, I * it is, is not * for, at the * time, to * in the, is a * of, I think * should, is very * to, the * and the, in the * and, if * want to, of the * and, a * of people, to the * and, the * is not, have a * of*
Advanced	13	*it is * to, I * it is, I think * is, have the * to, to * in the, I think * should, is a * of, the * and the, to the * and, in the * and, if * want to, to * to the, have to * the*
Native English Speakers	15	*the * of the, I think * is, and I * that, should be * to, they are * to, to be * to, would be * to, to * in the, a * of the, I would * to, in the * and, the * and the, a good * for, to * able to, as * as they*

We also compared the mean variability of the p-frames used in each group's written and oral production data. As shown in Table 14, the beginner, low-intermediate, and NES groups used significantly more variable p-frames in oral than written data, whereas the advanced group used significantly more variable p-frames in writing than in speech. The high-intermediate group showed no differences between written and oral data.

Table 14. Variability of P-frames across Written and Oral Datasets

Group	Mean Variability (SD) in Writing	Mean Variability (SD) in Speaking	Wilcoxon W	p-value
Beginner	0.29 (0.24)	0.40 (0.26)	6424.00	.000***
Low-intermediate	0.22 (0.19)	0.37 (0.24)	6912.00	.000***
High-intermediate	0.32 (0.23)	0.32 (0.22)	5074.00	.857
Advanced	0.44 (0.27)	0.34 (0.24)	3873.00	.006**
Native English Speakers	0.38 (0.25)	0.56 (0.27)	6964.50	.000***

**$p < .01$, ***$p < .001$

To determine other potential differences in learners' use of p-frames between written and oral production data, we also took a more qualitative approach and looked more closely at items that were used by all groups in both written and oral modalities. As shown in Table 15, there were a total of four

p-frames that qualified for this part of the analysis: *I think * is, to * in the, in the * and,* and *the * and the*. The top three variants for each p-frame and dataset are also provided.

Table 15. The Top Three Variants of P-frames Used in All Written and Oral Datasets

Group	Mode	*I think * is*	*to * in the*	*in the * and*	*the * and the*
Beginner	Written	it, this, smoking	smoke, work, do	restaurant, restaurants, country	smoke, study, school
	Oral	it, she, speaking	swim, play, live	park, sea, beach	woman, mother, smokes
Low-intermediate	Written	it, smoking, this	smoke, work, do	restaurant, restaurants, future	smoker, restaurant, smoke
	Oral	it, she, this	smoke, swim, play	sea, beach, park	mother, park, study
High-intermediate	Written	it, this, smoking	smoke, work, do	restaurant, future, country	smoker, smokers, user
	Oral	it, she, speaking	swim, play, smoke	park, restaurant, sea	mother, smoke, park
Advanced	Written	it, that, this	work, smoke, be	country, streets, society	smokers, smoker, restaurant
	Oral	it, this, she	swim, smoke, play	park, restaurant, sea	restaurant, windows, park
Native English Speakers	Written	it, this, that	smoke, be, take	world, UK, restaurant	restaurants, right, environment
	Oral	it, she, that	smoke, play, be	park, past, water	smokers, atmosphere, situations

For three of the four p-frames (*I think * is, to * in the, in the * and*), we did not find any major differences across levels and modes in the use of most common variants. However, for the p-frame *the * and the*, we observed an interesting difference in the use of inflected and derived forms of "smoke" by different learner groups Specifically, the beginner learners used the noun "smoke(s)" in their written and oral production. However, as proficiency increased, learners started to also attach the derivational suffix *–er* to the noun refer to a person who smokes (i.e., smoker(s)). This was observed only in the learners' written production and not in their oral production.

5. Discussion

The current study explored three research questions regarding the use of p-frames by L2 learners. Specifically, it investigated the p-frames in learners' written and oral productions using the WE and SD sub-corpora of ICNALE, which were similar in terms of topics and prompts. The first research question aimed to examine the variability, predictability, and discourse functions of the most frequently occurring p-frames in L2 written production. The results showed that the L2 learners used more variable p-frames in their writing as their proficiency level increased. Also, the predictability of selected p-frames became closer to that of the L1 reference group as proficiency increased. These findings align with those in previous studies that examined the developmental trajectories of the variability and predictability of p-frames in L2 writing, using cross-sectional data (Garner, 2016; Tan & Römer, 2022).

However, in terms of the discourse functions of the p-frames in the written data, the current study found no statistically significant developmental patterns. This result contrasts with Tan and Römer (2022), in which the use of stance expressions increased significantly as proficiency increased. Garner (2016) also observed that the discourse functions of learners' p-frames diversified as proficiency

increased. Specifically, writers at high-intermediate and advanced levels used more special conversational expressions and discourse organizing expressions than writers at the beginner level, demonstrating learners' ability to use p-frames to fulfill a wider variety of discourse functions as their proficiency level increased. The current study did not find such diversification of discourse functions as proficiency level increased. The p-frames included in our analysis mostly functioned as stance expressions or referential expressions and not many discourse organizers and special conversational expressions appeared in the data. The current study's absence of developmental patterns in the discourse functions of p-frames could be explained by the nature of the tasks and prompts that the learners were asked to complete. The writing tasks in EFCAMDAT, which were analyzed in Garner (2016), and Tan and Römer (2022), varied across proficiency levels. For instance, learners at the lower levels were asked to complete more simple, informal, and descriptive tasks (e.g., introducing yourself, describing your favorite day), while learners at the higher levels were asked to complete more complicated, formal, and opinion-giving tasks (e.g., giving advice about budgeting, writing a movie review). The WE sub-corpus used in the current study, however, used the same tasks for all groups, which likely contributed to the limited functional diversification observed in our data. Although the results were not statistically significant, we observed that the beginner learners used more stance expressions than referential expressions, in contrast to the other learner groups (i.e., low-intermediate, high-intermediate, advanced) and the L1 reference group, who used more referential expressions. Focusing on only the business email data in EFCAMDAT, Xia et al. (2023) also found that learners used significantly more stance expressions than referential expressions compared to the working professionals. The interpretation drawn from this result was that the learners heavily relied on certain stance frames with high frequency to express their opinions, attitude, or intention (e.g., *I would * to, have been * to, we can * a*). However, this interpretation needs to be treated with caution, as Xia et al. (2023) did not look at the developmental patterns across different L2 proficiency levels but combined the intermediate and advanced level learners' production data and compared them with working professionals' production data.

Our second research question aimed to investigate the variability, predictability, and discourse functions of the most frequently used p-frames in L2 learners' oral production data. In contrast to the written data, no developmental patterns were found in the variability of p-frames. Rather, all learner groups used significantly more fixed p-frames than the L1 reference group. This finding contrasts with Römer and Banerjee (2017) and Nekrasova-Beker (2021), which both revealed a developmental trajectory in learners' productivity of p-frames in speaking as they moved up to a higher language proficiency. Also, we were not able to find a clear developmental pattern of predictability in the selected p-frames produced by learners, as the learners did not demonstrate a greater alignment with the L1 reference group as their proficiency increased. This finding contrasts with the results discussed in Römer and Garner (2019), which observed that the set of verbs in the target constructions was more predictable and closer to the pattern exhibited by the L1 reference group among higher proficiency learners with Romance language backgrounds (e.g., Italian and Spanish) as opposed to those with lower proficiency levels. Based on the current study's findings, we could cautiously conclude that for Asian learners of English, speaking presents a greater challenge in terms of using variable multi-word sequences and aligning with the L1 reference group than writing does.

Additionally, the current study did not observe any significant developmental patterns in the discourse functions of p-frames produced in the spoken data. Similar to the findings for the written data, the absence of such patterns could be due to task or prompt effects, as the learners in the SD sub-corpus were told to (1) describe a series of pictures related to part-time jobs and to (2) persuade somebody to agree with their opinions in a role-play task. The use of stance expressions is expected, especially in the second task. The collaborative oral tasks used in Nekrasova-Beker (2021) included a wider range of prompts and tasks (e.g., persuasion, decision making, selecting from alternatives). Also, the MET speaking test examined in Römer and Banerjee (2017) consisted of tasks that involved various

communicative functions (e.g., describing a picture, talking about a personal experience, giving a personal opinion, explaining the advantages and disadvantages of an option, and persuading somebody). Thus, the focus on a few tasks in the SD sub-corpus could have led to less variation in the use of functions across groups. The frequency data in the current study, however, revealed that all learner groups predominantly used stance expressions in their oral productions, in contrast to the L1 reference group, which used more referential expressions. Biber et al. (2004) reported that referential expressions are more prevalent in academic writing, while stance expressions are more commonly used in conversations. Hence, a possible interpretation of the learners' greater usage of stance expressions in the current study could be that their utterances resembled the conversational register more closely than academic texts, whereas the utterances by the L1 reference group more closely resembled academic English.

Finally, our last research question addressed the differences in L2 learners' use of p-frames between their written and oral production. To answer this research question, both quantitative and qualitative analyses were conducted. First, there was a gradual decrease in the number of overlapping p-frames between the written and oral production data as proficiency level increased. To be specific, lower proficiency learners (i.e., beginner and low-intermediate groups) had a greater number of overlapping p-frames between written and oral datasets than high proficiency learners (i.e., advanced group) and the L1 reference group. This means that learners with lower proficiency showed fewer register differences between writing and speaking than learners of higher proficiency and native English speakers. Second, a comparison of the variability of p-frames between written and oral data revealed that beginner and low-intermediate groups used significantly more variable p-frames in oral production, similar to the native English speaker group. Advanced learners, however, used significantly more variable p-frames in their written productions. The high-intermediate group did not differ significantly in the two modalities. This could mean that learners at low proficiency levels are more willing to take risks in using variable frames in speaking than in writing. However, as they become more proficient in their L2, they incorporate variable sequences of words more easily in their writing than in their speaking. Advanced learners have likely received more language input including a wider range of p-frame realizations and appear to be more confident in incorporating what they have learned or been exposed to in writing than speaking. For example, the p-frame *is a * of* appeared in advanced learners' written and oral production data. The learners used a wide range of variants in this p-frame in writing (e.g., *waste, benefit, place, danger*, and *demerit*) but relied heavily on the "phrasal teddy bear" (Ellis, 2012) *is a lot of* in their oral data. Dyadic oral tasks can be challenging for learners and don't allow for much planning time. There are also affective factors (e.g., anxiety, confidence) which may influence learners' spoken language production (Cheng, Horwitz, & Schallert, 1999; Horwitz, Horwitz, & Cope, 1986) and the spontaneity of the mode (Boers et al., 2006). However, it is important to note one caveat in this analysis. The low proficiency learners who contributed to our datasets occasionally produced inaccurate variants in their oral production of p-frames. For instance, when examining the variants of the p-frame *want to * a,* the written data predominantly featured accurate basic verb forms (e.g., *be, have, do, get,* and *try*), while the oral data contained inaccurate verb forms, such as instances of *did.* Since we did not exclude these inaccurate variants in our dataset, this could have mildly affected the higher mean variability of p-frames in the low-level learners' oral data. Finally, the third qualitative analysis we conducted by further zooming in on the data showed that learners potentially found it easier to try new morphemes (e.g., *-er* in *smoker*) in writing than in speaking. Similar to the use of different word types in the slots of p-frames, this finding confirms our observation that learners are able to incorporate a greater variety of p-frame realizations in their writing than in their speaking.

We think that the findings of our study have relevant implications for SLA theory and L2 pedagogy. The results suggest that using p-frames as a unit of analysis may be a suitable measure to examine differences in language patterns across different L2 proficiency levels and modalities (i.e., written vs.

oral). Further, the results pertaining to the third research question, which complement existing literature on p-frame use by L2 learners, provide implications that there exist disparities between learners' written and oral production. Specifically, achieving the target norm in speaking with respect to this phraseological item may be a more challenging productive skill compared to writing. In terms of pedagogical applications, while the findings may not have direct implications for L2 teaching, they can offer valuable insights to English as an L2 practitioners in Asian countries in which students often lack exposure to authentic English in and outside of classroom settings. To be specific, practitioners could consider incorporating more high-frequency p-frames collected from L1 reference data in L2 materials and instruction. Recently, artificial intelligence (AI)-powered tools, such as ChatGPT, have become popular resources to support teachers in creating data-driven L2 learning materials (Mizumoto, 2023). Teachers could utilize these AI platforms to generate texts that incorporate p-frames that are frequently used in the target language and supplement existing teaching materials with those texts. Teachers could then develop consciousness-raising tasks that enable learners to read an AI-generated text that contains target-language p-frames and analyze the text by identifying recurring patterns, such as different realizations of frequently occurring 4-word frames (e.g., *I do think that, I do know that, I do believe that*). Based on our findings, which highlight differences in the use of p-frames between speech and writing, practitioners are encouraged to tailor the selection of examples to the modality the lesson is focusing on.

6. Conclusions

The current study investigated L2 learners' use of p-frames in both written and oral production across proficiency levels. The results revealed that in written production, learners used more variable p-frames as their proficiency increased. Results also indicated that the predictability of p-frames produced by learners became more similar to the L1 reference group as learners became more proficient. There were no clear developmental trajectories for discourse functions of p-frames, the distributions of which remained stable across levels. In the oral data, all learner groups used more fixed p-frames than the L1 reference group. Different from what we observed for learner writing, the predictability of p-frames did not increase in speech as proficiency increased. The distribution of p-frames across function categories did not change significantly from lower to higher proficiency levels either. A comparison of learners' use of p-frames in written and oral production indicated that the lower-level learners shared more common p-frames across the two modes than the advanced learner and L1 reference groups. Additionally, the beginner and low-intermediate groups and the L1 reference group used more variable sequences in their writing, whereas the advanced learners used more variable sequences in their speaking. The findings of our study provide pedagogically relevant insights for English as an L2 practitioners in Asian countries. More specifically, we believe that practitioners could utilize some of the high-frequency p-frames and their typical variants extracted from L1 reference data to create new or improve existing L2 teaching materials.

Our study has several limitations that ought to be addressed in future work on the topic. First, learners with different L1 backgrounds were grouped together in the current study to give us sufficiently robust word counts in all sub-corpora to be able to conduct statistical analyses. Although Asian languages are very diverse, it can be argued that some of them are typologically related (e.g., Japanese and Korean; Phuoc & Barrot, 2022). Future research, however, could divide learner texts further into L1 groups to enable a more nuanced analysis of potential differences between learner groups and examine the role of the learners' first languages on their use of p-frames. Another limitation is that, due to the exclusion of the interviewers' utterances, the oral sub-corpus used in the current study was smaller than the written sub-corpus. To allow for a more accurate comparison of p-frames between the two modes, future studies could attempt to more carefully balance written and

oral sub-corpora in terms of size. Third, in our analysis of discourse functions expressed by the most common p-frames in each dataset, we investigated a limited number of general functional categories discussed in previous research (i.e., stance expression, referential expression, discourse organizer, and special conversational expression). Assigning one of only four broad categories to each p-frame can pose limitations on fully capturing the more specific function that each p-frame conveys. Thus, a more fine-grained analysis of the functions of the p-frame would be necessary. For example, future research could code the subcategories for each discourse function (e.g., breaking down stance expressions into (1) attitudinal/modality expressions that overtly show the writer's attitude, as in *I think * is,* and (2) anticipatory *it* expressions, as in *it is * to*; Chen & Baker, 2016). Fourth, the data examined in the current study were limited to mostly the persuasive genre, as the prompts in ICNALE required the participants to agree or disagree with the topics (WE sub-corpus) and persuade another person in specific situations (SD sub-corpus). While focusing on one type of genre is valuable for understanding learners' repertoire of discourse functions within that specific context, exploring the use of p-frames in a wider range of genres (e.g., narratives, descriptions) would provide a more comprehensive view of L2 learners' phraseological competence and its development. Finally, the current study examined only the top-100 4-word p-frames in each sub-corpus. It would be helpful if future studies examined longer results lists and p-frames of different lengths (e.g., 3- and 5-word frames). Despite these limitations, we think that our study makes an important contribution to the growing body of research on the significance of phraseology in conceptualizing L2 development. We hope to see future studies examining learners' use of p-frames in additional longitudinal and cross-sectional corpora (both written and oral) to further enhance our understanding of the development of L2 phraseological competence.

Acknowledgments

This study was first presented at the American Association of Applied Linguistics (AAAL) 2023 Conference in Portland, OR. We express our gratitude to audience members who provided invaluable comments on our work.

References

Anthony, L. (2022). AntConc (Version 4.0.5). [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

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.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263-286.

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

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research, 10*(3), 245-261.

Chan, H., Verspoor, M., & Vahtrick, L. (2015). Dynamic development in speaking versus writing in identical twins. *Language Learning, 65*(2), 298-325.

Chen, Y. H., & Baker, P. (2016). Investigating critical discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2, and C1. *Applied Linguistics, 37*(6), 849-880.

Cheng, Y., Horwitz, E. K., & Schallert, D. L. (1999). Language anxiety: Differentiating writing and speaking components. *Language Learning, 49*(3), 417-446.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics, 32*, 17-44.

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), 31-68.

Gilquin, G. (2015). The use of phrasal verbs by French-speaking EFL learners: A constructional and collostructional corpus-based approach. *Corpus Linguistics and Linguistic Theory, 11*, 51-88.

Goldberg, A. (2013). Constructionist approaches. In Hoffmann, T., & Trousdale, G. (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 15-31). Oxford: Oxford University Press.

Gray, B., & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics, 18*(1), 109-136.

Gries, S. T., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning 65*(S1), 228-255.

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal, 70*(2), 125-132.

Hwang, H., Jung, H., & Kim, H. (2020). Effects of written versus spoken production modalities on syntactic complexity measures in beginning-level child EFL learners. *The Modern Language Journal, 104*(1), 267-283.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and the World, 1,* 91-118.

Ishikawa, S. (2019). The ICNALE spoken dialogue: A new dataset for the study of Asian learners' performance in L2 English interviews. *English Teaching (The Korea Association of Teachers of English), 74*(4), 153-177.

Ishikawa, S. (2023). *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*. New York, NY: Routledge.

Juknevičienė, R., & Grabowski, Ł. (2018). Comparing formulaicity of learner writing through phrase-frames: A corpus-driven study of Lithuanian and Polish EFL student writing. *Research in Language, 16*(3), 303-323.

Kassambara, A. (n.d.). Kruskal-Wallis effect size. Retrieved April 26, 2022, from https://rpkgs.datanovia.com/rstatix/reference/kruskal_effsize.html

Kumar, U., Kumar, V., & Kapur, J. N. (1986). Normalized measures of entropy. *International Journal of General Systems, 12*(1), 55-69.

Larsson, T., Reppen, R., & Dixon, T. (2022). A phraseological study of highlighting strategies in novice and expert writing. *Journal of English for Academic Purposes, 60,* 101179.

Liu, L., Jiang, F., Du, Z. (2023). Figure legends of scientific research articles: Rhetorical moves and phrase frames. *English for Specific Purposes, 70,* 86-100.

Lu, X., Yoon, J., & Kisselev, O. (2018). A phrase-frame list for social science research article introductions. *Journal of English for Academic Purposes, 36,* 76-85.

Mizumoto, A. (2023). Data-driven learning meets generative AI: Introducing the framework of metacognitive resource use. *Applied Corpus Linguistics, 3,* 100074.

Nekrasova-Beker, T. (2009). English L1 and L2 speakers' knowledge of lexical bundles. *Language Learning, 59*(3), 647-686.

Nekrasova-Beker, T. (2021). Use of phrase-frames in L2 students' oral production across proficiency sub-levels. Crawford, W. J. (Ed.), *Multiple Perspectives on Learner Interaction: The Corpus of Collaborative Oral Tasks* (pp. 41-68). Berlin: De Gruyter Mouton.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

O'Donnell. M. B., Römer, U., & Ellis, N. (2013). The development of formulaic sequences in first and second language writing. *International Journal of Corpus Linguistics, 18*(1), 83-108.

Paquot, M. (2013). Lexical bundles and L1 transfer effects. *International Journal of Corpus Linguistics, 18*(3). 391-417.

Paquot, M. (2014). Cross-linguistic influence and formulaic language: Recurrent word sequences in French learner writing. *EUROSLA Yearbook, 14,* 240-261.

Phuoc. V. D., & Barrot, J. S. (2022). Complexity, accuracy, and fluency in L2 writing across proficiency levels: A matter of L1 background? *Assessing Writing, 54*, 100673.

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics, 7,* 140-162.

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.

Römer, U., & Banerjee, J. (2017). Validating the MET speaking test through phraseological analysis: A corpus approach to language assessment. *CaMLA Working Papers, 2017-01*, 1-26.

Römer, U., & Garner, J. R. (2019). The development of verb constructions in spoken learner English: Tracking effects of usage and proficiency. *International Journal of Learner Corpus Research, 5*(2), 207-230.

Tan, Y., & Römer, U. (2022). Using phrase-frames to trace the language development of L1 Chinese learners of English. *System, 108,* 1-10.

Wulff, S. (2019). Acquisition of formulaic language from a usage-based perspective. In Siyanova-Chanturia, A., & Pellicer-Sánchez, A. (Eds.), *Understanding Formulaic Language: A Second Language Acquisition Perspective* (pp. 19-37). New York, NY: Routledge.

Xia, D., Sulzer, M. A., & Pae, H. K. (2023). Phrase-frames in business emails: A contrast between learners of business English and working professionals. *Text & Talk, 1-22*.

Appendix

Appendix 1. The Top 100 P-frames in Written Production (Beginner)

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
1	I think * is	59.72	0.09	0.41
2	it is * to	57.29	0.33	0.83
3	I * it is	50.35	0.07	0.26
4	the * of the	49.31	0.75	0.96
5	that * should be	41.32	0.04	0.18
6	I agree * the	33.68	0.07	0.36
7	think that * is	33.68	0.17	0.48
8	I think * should	32.29	0.15	0.63
9	I * that it	31.25	0.11	0.42
10	I * with the	30.21	0.05	0.22
11	that * is important	27.08	0.06	0.17
12	is very * for	26.74	0.25	0.76
13	is not * for	26.04	0.27	0.71
14	it is * that	26.04	0.44	0.90
15	to * a part	25.69	0.12	0.61
16	to * in the	25.69	0.37	0.83
17	agree with * statement	25.69	0.04	0.64
18	that it * important	25.69	0.01	0.00
19	think * it is	23.96	0.03	0.11
20	I think * it	23.96	0.04	0.19
21	I agree * this	23.26	0.06	0.33
22	think * is important	22.57	0.03	0.12
23	I * with this	22.22	0.03	0.40
24	a good * to	22.22	0.34	0.63
25	there are * reasons	22.22	0.13	0.78
26	think it * important	22.22	0.02	0.00
27	think that * should	22.22	0.14	0.64
28	a * of money	21.88	0.05	0.15
29	if * want to	20.14	0.16	0.73
30	it is * good	19.79	0.14	0.70
31	is not * to	19.44	0.50	0.89
32	a * of people	18.75	0.07	0.20
33	think * should be	18.75	0.13	0.34
34	a lot * people	18.40	0.04	0.23
35	in the * and	18.40	0.53	0.92
36	I have * reasons	18.06	0.12	0.76
37	is very * to	18.06	0.29	0.89
38	you * do it	17.71	0.10	0.86
39	in the * of	17.71	0.78	0.96
40	if * have a	17.36	0.10	0.85
41	is * important for	17.36	0.20	0.72
42	the * of money	17.01	0.22	0.66
43	want to * a	17.01	0.35	0.77
44	is a * of	16.67	0.50	0.86
45	the * is not	16.32	0.68	0.92
46	have a * of	16.32	0.23	0.42
47	can * a lot	15.97	0.50	0.82
48	with the * that	15.28	0.16	0.75
49	it is * important	14.93	0.23	0.77
50	a good * for	14.58	0.48	0.90
51	not only * but	14.58	0.67	0.96
52	the * should be	14.24	0.56	0.90
53	for the * of	14.24	0.59	0.93

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
54	on the * hand	14.24	0.05	0.66
55	will be * to	14.24	0.42	0.64
56	is * good for	13.89	0.08	0.40
57	to * a good	13.89	0.28	0.76
58	to * a lot	13.89	0.50	0.94
59	to the * of	13.89	0.85	0.97
60	have * lot of	13.54	0.05	0.29
61	do not * to	13.89	0.33	0.74
62	in the * is	13.54	0.28	0.77
63	a * way to	12.85	0.05	0.18
64	agree with * opinion	12.85	0.11	0.69
65	I * agree with	12.50	0.36	0.72
66	the * and the	12.50	0.86	0.98
67	are a * of	12.50	0.22	0.59
68	only * but also	12.15	0.66	0.94
69	important for * to	12.15	0.20	0.75
70	people who * to	12.15	0.34	0.70
71	I * that the	11.81	0.24	0.64
72	is * a good	11.81	0.21	0.53
73	on * other hand	11.81	0.03	0.00
74	that * is not	11.81	0.24	0.65
75	the * of a	11.81	0.77	0.97
76	in the * because	11.81	0.12	0.64
77	is very * and	11.81	0.53	0.92
78	people who * not	11.81	0.15	0.43
79	should be * in	11.81	0.29	0.54
80	and * in the	11.11	0.85	0.97
81	be * at all	11.46	0.12	0.36
82	the * reason is	11.46	0.21	0.78
83	is not * good	11.46	0.18	0.52
84	know the * of	11.46	0.39	0.78
85	is * for us	11.11	0.44	0.77
86	is * good way	11.11	0.09	0.34
87	for a * time	11.11	0.06	0.93
88	I think * we	11.11	0.16	0.40
89	it is * a	11.11	0.38	0.59
90	there are * people	11.11	0.28	0.68
91	do not * the	10.76	0.48	0.88
92	in the * place	10.76	0.32	0.79
93	is a * to	10.76	0.48	0.82
94	matter * you are	10.42	0.10	0.95
95	not * in the	10.42	0.50	0.78
96	and the * of	10.42	0.97	1.00
97	is a * way	10.42	0.07	0.21
98	is the * of	10.42	0.83	0.98
99	of * in the	10.07	0.72	0.93
100	think * is a	10.07	0.21	0.77

Appendix 2. The Top 100 P-frames in Written Production (Low-intermediate)

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
1	the * of the	53.43	0.59	0.94
2	it is * to	49.19	0.32	0.79
3	that * should be	43.76	0.04	0.15
4	I think * is	42.91	0.07	0.35
5	I * it is	36.98	0.04	0.21
6	it is * that	30.53	0.41	0.86

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
7	is very * for	30.36	0.16	0.69
8	is not * for	28.50	0.13	0.66
9	I think * should	27.65	0.14	0.58
10	on the * hand	25.61	0.03	0.38
11	there are * reasons	24.09	0.05	0.83
12	should be * in	22.90	0.17	0.40
13	a good * to	22.56	0.23	0.72
14	to * in the	22.05	0.32	0.81
15	I * with the	21.71	0.02	0.57
16	that * is important	21.71	0.03	0.10
17	is * important for	21.37	0.11	0.63
18	I agree * the	21.20	0.07	0.26
19	that it * important	21.20	0.01	0.00
20	have a * of	21.03	0.13	0.34
21	it is * good	21.03	0.11	0.60
22	I * that it	20.35	0.09	0.47
23	think that * is	20.18	0.10	0.38
24	a * of money	19.51	0.06	0.23
25	think * should be	18.83	0.10	0.40
26	agree with * statement	19.00	0.04	0.71
27	is * good for	17.81	0.11	0.45
28	is a * of	17.81	0.41	0.84
29	have * lot of	17.13	0.01	0.00
30	I * with this	17.13	0.02	0.53
31	in the * of	16.96	0.67	0.96
32	a * of people	16.45	0.08	0.25
33	I agree * this	16.45	0.05	0.24
34	if * have a	16.28	0.09	0.77
35	be * at all	16.11	0.10	0.30
36	if * want to	16.11	0.10	0.66
37	think * it is	15.61	0.02	0.09
38	for the * of	15.61	0.40	0.82
39	with the * that	15.61	0.23	0.74
40	we can * a	15.10	0.35	0.87
41	think that * should	14.93	0.17	0.52
42	a lot * people	14.76	0.01	0.00
43	think * is important	14.59	0.04	0.16
44	will be * to	14.59	0.41	0.74
45	the * reason is	14.42	0.14	0.76
46	I think * it	14.08	0.04	0.12
47	they can * their	14.08	0.51	0.89
48	can * a lot	13.91	0.38	0.82
49	it is * important	13.91	0.15	0.66
50	at the * time	13.74	0.06	0.23
51	in the * and	13.74	0.43	0.91
52	in the * is	13.74	0.24	0.73
53	is not * to	13.57	0.46	0.83
54	people who * not	13.57	0.09	0.55
55	how to * with	13.40	0.28	0.76
56	is very * to	13.40	0.29	0.83
57	want to * a	13.23	0.23	0.77
58	have the * to	13.06	0.26	0.71
59	the * and the	12.72	0.69	0.96
60	as we * know	12.72	0.03	0.10
61	are a * of	12.55	0.14	0.34
62	all * in the	12.38	0.15	0.38
63	the * is not	12.21	0.58	0.89

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
64	and the * of	12.21	0.75	0.96
65	there * a lot	12.04	0.07	0.38
66	there are * people	12.04	0.20	0.56
67	agree that * should	11.87	0.07	0.19
68	in the * place	11.70	0.22	0.69
69	it is * a	11.53	0.21	0.69
70	lot of * to	11.53	0.46	0.80
71	not only * the	11.53	0.47	0.83
72	a * of time	11.36	0.09	0.60
73	is a * way	11.36	0.19	0.42
74	the * should be	11.20	0.36	0.82
75	we * have a	11.03	0.17	0.72
76	a * job in	10.86	0.13	0.44
77	have * right to	10.86	0.11	0.68
78	have to * a	10.86	0.41	0.86
79	are * lot of	10.69	0.03	0.12
80	is * of the	10.69	0.16	0.36
81	the * of their	10.69	0.70	0.93
82	a good * for	10.69	0.37	0.88
83	do not * to	10.69	0.22	0.68
84	be * in the	10.52	0.45	0.77
85	if you * to	10.52	0.18	0.48
86	is the * of	10.52	0.65	0.95
87	of * in the	10.35	0.49	0.84
88	the * important thing	10.35	0.08	0.21
89	if you * a	10.35	0.21	0.62
90	there are * lot	10.35	0.03	0.12
91	the * will be	10.18	0.63	0.94
92	but also * the	10.18	0.43	0.80
93	not only * but	10.18	0.45	0.90
94	the most * thing	10.18	0.05	0.21
95	have * time to	10.01	0.27	0.84
96	I * agree with	10.01	0.29	0.84
97	is * good way	10.01	0.07	0.35
98	that * is not	10.01	0.14	0.57
99	it * a good	9.84	0.12	0.33
100	have a * to	9.84	0.43	0.78

Appendix 3. The Top 100 P-frames in Written Production (High-intermediate)

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
1	the * of the	78.21	0.50	0.90
2	that * should be	50.22	0.04	0.21
3	it is * to	41.10	0.36	0.84
4	should be * in	31.67	0.13	0.34
5	having a * time	29.91	0.03	0.11
6	I think * is	25.43	0.10	0.37
7	in the * of	24.47	0.55	0.93
8	it is * that	23.99	0.37	0.88
9	for the * of	22.39	0.44	0.85
10	I * it is	22.07	0.06	0.24
11	is not * for	21.91	0.22	0.65
12	be * at all	21.11	0.11	0.27
13	is * important for	20.79	0.16	0.68
14	I * that it	19.35	0.15	0.65
15	that * is important	19.03	0.03	0.12
16	agree that * should	17.91	0.05	0.20

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
17	at the * time	17.91	0.03	0.09
18	on the * hand	17.91	0.02	0.52
19	to the * of	17.75	0.71	0.93
20	it is * important	17.59	0.14	0.68
21	not only * the	17.11	0.47	0.88
22	is very * for	16.79	0.20	0.71
23	to * in the	16.47	0.42	0.86
24	the * of a	15.99	0.77	0.97
25	is a * of	15.83	0.36	0.84
26	the * of their	14.40	0.62	0.94
27	I think * should	14.40	0.18	0.65
28	with the * of	14.40	0.71	0.96
29	is * of the	14.24	0.11	0.33
30	have the * to	14.08	0.30	0.75
31	is not * to	14.24	0.55	0.92
32	is the * of	13.92	0.66	0.95
33	is very * to	13.76	0.27	0.85
34	think that * is	13.76	0.09	0.49
35	to be * to	13.44	0.43	0.69
36	the * and the	13.28	0.63	0.93
37	there are * reasons	13.12	0.17	0.79
38	a good * to	12.80	0.26	0.73
39	in the * and	12.80	0.48	0.90
40	they can * their	12.64	0.51	0.89
41	a * of money	12.48	0.13	0.34
42	of the * of	12.48	0.86	0.99
43	I * with the	12.16	0.04	0.57
44	is one * the	12.16	0.03	0.10
45	that * is not	11.68	0.11	0.55
46	with the * that	11.68	0.33	0.72
47	on the * of	11.52	0.72	0.95
48	they are * to	11.52	0.49	0.84
49	will be * to	11.52	0.42	0.75
50	it is * a	11.36	0.30	0.72
51	if * want to	11.20	0.17	0.65
52	and the * of	11.04	0.71	0.96
53	agree with * statement	10.88	0.06	0.59
54	it is * good	10.72	0.10	0.53
55	so that * can	10.72	0.13	0.72
56	do not * to	10.56	0.30	0.80
57	know that * is	10.56	0.15	0.55
58	I agree * the	10.40	0.09	0.34
59	be * in the	10.24	0.58	0.86
60	a good * for	10.24	0.48	0.90
61	in the * is	10.24	0.31	0.80
62	is * good for	10.08	0.11	0.42
63	people who * not	10.08	0.11	0.51
64	it * be a	9.92	0.18	0.87
65	of * in the	9.92	0.61	0.91
66	but also * the	9.92	0.44	0.81
67	of the * and	9.92	0.63	0.94
68	think * it is	9.76	0.03	0.12
69	a * of people	9.44	0.10	0.35
70	the * that they	9.44	0.63	0.95
71	to * their time	9.44	0.25	0.75
72	I strongly * that	9.44	0.17	0.63
73	the * who are	9.28	0.33	0.68

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
74	the * of money	9.12	0.23	0.62
75	think * should be	9.12	0.11	0.42
76	for * to have	8.96	0.18	0.74
77	as a * of	9.12	0.48	0.87
78	in the * place	8.96	0.20	0.71
79	to be * in	8.96	0.77	0.96
80	a * of time	8.80	0.13	0.64
81	I * agree with	8.80	0.36	0.84
82	the * but also	8.80	0.38	0.78
83	do not * the	8.80	0.49	0.85
84	the most * thing	8.80	0.13	0.31
85	to the * and	8.80	0.69	0.93
86	have * right to	8.64	0.09	0.71
87	I * agree that	8.64	0.32	0.71
88	the * reason is	8.64	0.19	0.81
89	to * in a	8.48	0.45	0.86
90	is a * that	8.48	0.64	0.92
91	in the * because	8.16	0.23	0.65
92	there are * many	8.32	0.14	0.70
93	the * effects of	8.16	0.29	0.77
94	the * of people	8.16	0.55	0.86
95	would be * to	8.16	0.55	0.88
96	not be * to	8.00	0.44	0.72
97	the * is not	7.84	0.65	0.96
98	people * do not	7.84	0.08	0.35
99	have a * of	7.84	0.27	0.55
100	a * in the	7.68	0.69	0.90

Appendix 4. The Top 100 P-frames in Written Production (Advanced)

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
1	that * should be	68.14	0.08	0.25
2	it is * to	48.40	0.51	0.88
3	that * is important	35.03	0.06	0.22
4	it is * that	34.39	0.50	0.92
5	I * that it	33.75	0.15	0.87
6	that it * important	33.12	0.02	0.00
7	I * it is	31.84	0.12	0.38
8	I think * is	31.21	0.10	0.37
9	have a * time	30.57	0.06	0.42
10	should be * in	29.93	0.23	0.53
11	have the * to	29.30	0.33	0.71
12	for the * of	23.56	0.54	0.89
13	is very * for	22.29	0.43	0.81
14	I * with the	21.65	0.06	0.67
15	in the * of	21.02	0.73	0.97
16	have * right to	20.38	0.25	0.54
17	is * important for	20.38	0.25	0.78
18	agree with * statement	20.38	0.09	0.66
19	is not * for	20.38	0.38	0.80
20	with the * that	20.38	0.31	0.70
21	the * of a	19.74	0.81	0.98
22	agree that * should	19.74	0.10	0.26
23	I agree * the	19.74	0.10	0.35
24	it is * important	19.74	0.23	0.83
25	to the * of	19.74	0.90	0.99
26	to * in the	18.47	0.52	0.88

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
27	at the * time	18.47	0.17	0.43
28	would be * to	18.47	0.52	0.86
29	of the * of	17.83	0.89	0.98
30	is * of the	17.20	0.07	0.38
31	the * of money	16.56	0.23	0.66
32	should be * to	16.56	0.54	0.92
33	I * agree with	15.92	0.36	0.81
34	is one * the	15.92	0.04	0.00
35	there are * reasons	15.92	0.20	1.00
36	to be * to	15.92	0.56	0.84
37	do not * to	14.65	0.39	0.81
38	will be * to	14.65	0.52	0.78
39	agree that * is	14.01	0.09	0.58
40	some people * that	14.01	0.32	0.88
41	they are * to	14.01	0.73	0.96
42	on * other hand	13.37	0.05	0.00
43	that * is not	13.37	0.29	0.73
44	the * should be	13.37	0.62	0.92
45	for the * to	13.37	0.38	0.89
46	I think * should	13.37	0.43	0.86
47	is * for the	12.74	0.60	0.95
48	is a * of	12.74	0.70	0.93
49	of the * and	12.74	0.85	0.98
50	for * to have	12.10	0.42	0.79
51	it * important to	12.10	0.11	0.30
52	the * and the	12.10	0.79	0.97
53	is more * than	12.10	0.47	0.88
54	is not * to	12.10	0.79	0.97
55	I * with this	11.46	0.11	0.85
56	the * of their	11.46	0.72	0.95
57	the * that it	11.46	0.44	0.77
58	to * their time	11.46	0.67	0.94
59	have to * a	11.46	0.67	0.95
60	I agree * it	11.46	0.11	0.65
61	is very * to	11.46	0.44	0.95
62	think that * is	11.46	0.33	0.65
63	to the * and	11.46	0.89	0.99
64	I * agree that	10.83	0.47	0.89
65	I * that the	10.83	0.24	0.89
66	the * will be	10.83	0.77	0.95
67	with * statement that	10.83	0.12	0.32
68	as a * of	10.83	0.53	0.90
69	in the * and	10.83	0.82	0.97
70	in the * is	10.83	0.41	0.85
71	in the * place	10.83	0.35	0.84
72	it is * good	10.83	0.29	0.90
73	not only * the	10.83	0.77	0.97
74	be * by the	10.19	0.88	0.97
75	be * in the	10.19	0.75	0.91
76	if * want to	10.19	0.44	0.88
77	it * be a	10.19	0.44	0.92
78	a good * for	10.19	0.56	0.88
79	and the * of	10.19	0.88	0.97
80	if the * is	10.19	0.56	0.86
81	should be * for	10.19	0.94	0.99
82	that the * of	10.19	0.88	0.99
83	a * of money	9.55	0.20	0.66

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
84	as * as possible	9.55	0.53	0.82
85	the * of people	9.55	0.40	0.89
86	be a * of	9.55	0.87	0.97
87	I have * reasons	9.55	0.40	0.90
88	on the * of	9.55	0.80	0.98
89	a * of time	8.92	0.21	0.87
90	the * of having	8.92	0.79	0.96
91	as * as they	8.92	0.43	0.86
92	there * be a	8.92	0.29	0.96
93	they * not have	8.92	0.21	0.73
94	to * to the	8.92	0.71	0.93
95	at the * of	8.92	0.50	0.89
96	do not * the	8.92	0.64	0.89
97	have to * the	8.92	0.93	0.99
98	I agree * this	8.92	0.14	0.37
99	important for * to	8.92	0.29	0.79
100	it is * a	8.92	0.43	0.86

Appendix 5. The Top 100 P-frames in Written Production (Native English Speakers)

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
1	I * that it	53.87	0.15	0.63
2	it is * to	53.87	0.43	0.93
3	the * of the	48.05	0.55	0.90
4	that * should be	44.41	0.20	0.58
5	I * think that	36.40	0.20	0.73
6	I think * is	36.40	0.12	0.63
7	think that * is	36.40	0.18	0.51
8	I don't * that	35.67	0.10	0.75
9	I think * it	29.85	0.05	0.17
10	and I * that	29.12	0.23	0.78
11	it is * important	29.12	0.28	0.78
12	should be * to	28.39	0.39	0.78
13	in the * of	26.94	0.62	0.91
14	think * it is	26.21	0.03	0.00
15	I * it is	24.75	0.09	0.32
16	do not * to	24.75	0.27	0.76
17	have the * to	24.75	0.32	0.66
18	I * believe that	23.30	0.31	0.75
19	the * of their	23.30	0.69	0.95
20	think that * should	23.30	0.28	0.71
21	should be * in	22.57	0.13	0.31
22	that * is important	21.84	0.10	0.35
23	the * of a	21.84	0.83	0.98
24	so that * can	21.84	0.17	0.75
25	if they * to	21.11	0.31	0.85
26	it is * that	21.11	0.72	0.96
27	that it * important	21.11	0.07	0.36
28	have * right to	19.66	0.07	0.83
29	that * is a	19.66	0.30	0.80
30	for the * of	19.66	0.63	0.91
31	is very * for	19.66	0.19	0.60
32	is not * to	18.20	0.64	0.90
33	that it * be	18.20	0.24	0.79
34	to the * of	18.20	0.80	0.93
35	I * that the	17.47	0.33	0.75
36	I * that they	17.47	0.46	0.86

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
37	is * of the	16.74	0.26	0.55
38	believe that * is	16.74	0.26	0.66
39	I do * that	16.74	0.30	0.84
40	they are * to	16.74	0.61	0.96
41	I * with this	16.02	0.09	0.98
42	that * would be	16.02	0.36	0.73
43	they * to do	16.02	0.23	0.82
44	and the * of	16.02	0.77	0.96
45	if * want to	15.29	0.33	0.81
46	is very * to	15.29	0.52	0.93
47	to be * to	15.29	0.29	0.58
48	would be * to	15.29	0.81	0.96
49	I * that this	14.56	0.40	0.86
50	the * that they	14.56	0.85	0.98
51	to * in the	14.56	0.80	0.97
52	have to * that	14.56	0.50	0.83
53	think that * would	14.56	0.40	0.77
54	a * of the	13.83	0.47	0.87
55	and * think that	13.83	0.16	0.37
56	a good * to	13.83	0.47	0.79
57	agree with * statement	13.83	0.16	0.74
58	I would * to	13.83	0.42	0.86
59	in the * and	13.83	0.63	0.92
60	of * in the	13.10	0.89	0.99
61	the * and the	13.10	0.94	0.99
62	it is * a	13.10	0.61	0.94
63	of the * and	13.10	1.00	1.00
64	as * result of	12.38	0.12	0.52
65	I * that I	12.38	0.53	0.89
66	it * be a	12.38	0.35	0.67
67	a good * for	12.38	0.53	0.93
68	as a * of	12.38	0.18	0.40
69	believe that * should	12.38	0.41	0.93
70	but I * that	12.38	0.29	0.70
71	is one * the	12.38	0.06	0.00
72	it is * good	12.38	0.24	0.48
73	that they * not	12.38	0.35	0.87
74	with the * of	12.38	1.00	1.00
75	a * way to	11.65	0.38	0.74
76	as * as possible	11.65	0.38	0.80
77	don't * that it	11.65	0.25	0.87
78	so * they can	11.65	0.06	0.00
79	the * is that	11.65	0.56	0.79
80	to * able to	11.65	0.13	0.34
81	to * in a	11.65	0.50	0.79
82	a large * of	11.65	0.44	0.96
83	at the * in	11.65	0.19	0.42
84	feel that * is	11.65	0.31	0.76
85	is a * way	11.65	0.13	0.81
86	to work * a	11.65	0.38	0.77
87	will be * to	11.65	0.56	0.91
88	a * of time	10.92	0.33	0.94
89	as * as they	10.92	0.33	0.88
90	do * want to	10.92	0.07	0.00
91	it * a good	10.92	0.13	0.35
92	should * able to	10.92	0.07	0.00
93	they * have to	10.92	0.60	0.95

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
94	think * is a	10.92	0.27	0.77
95	at the * time	10.92	0.13	0.35
96	in the * world	10.92	0.47	0.86
97	on the * of	10.92	0.87	0.98
98	that they * have	10.92	0.73	0.96
99	to do * they	10.92	0.40	0.89
100	with the * that	10.92	0.40	0.80

Appendix 6. The Top 100 P-frames in Oral Production (Beginner)

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
1	a few * ago	189.22	0.03	0.27
2	I think * is	130.44	0.26	0.75
3	it is * to	63.07	0.39	0.79
4	I * it is	51.60	0.14	0.31
5	I don't * to	51.60	0.17	0.72
6	so I * to	51.60	0.19	0.74
7	I * to go	48.74	0.21	0.73
8	I * go to	47.30	0.49	0.90
9	in the * and	47.30	0.55	0.88
10	I * like to	45.87	0.16	0.64
11	the * is very	40.14	0.75	0.98
12	want to * to	40.14	0.18	0.59
13	the * in the	38.70	0.93	0.99
14	to * to the	37.27	0.31	0.58
15	go to * park	37.27	0.12	0.49
16	when I * to	37.27	0.39	0.81
17	a * of people	35.84	0.08	0.24
18	the * of the	35.84	0.92	0.99
19	and I * to	35.84	0.48	0.87
20	with my * and	35.84	0.36	0.82
21	I * to the	34.40	0.25	0.64
22	I * want to	34.40	0.25	0.58
23	a lot * people	34.40	0.04	0.00
24	to the * and	34.40	0.54	0.94
25	and the * is	32.97	0.87	0.98
26	because I * to	32.97	0.26	0.83
27	I can * my	31.54	0.68	0.94
28	when I * a	31.54	0.23	0.52
29	I * to talk	30.10	0.14	0.66
30	the * and he	30.10	0.81	0.98
31	to * in the	30.10	0.48	0.89
32	I don't * so	30.10	0.29	0.58
33	he * to the	28.67	0.55	0.88
34	is * for me	27.24	0.53	0.94
35	so * want to	27.24	0.11	0.30
36	want * go to	27.24	0.05	0.00
37	I can * the	27.24	0.84	0.97
38	I would * to	27.24	0.16	0.37
39	is very * and	27.24	0.63	0.96
40	my * and I	25.80	0.78	0.93
41	have a * of	25.80	0.11	0.31
42	I want * go	25.80	0.06	0.00
43	to go * the	25.80	0.06	0.00
44	I * my friend	24.37	0.12	0.32
45	if * want to	24.37	0.24	0.79
46	when * was a	24.37	0.06	0.00

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
47	don't have * money	24.37	0.35	0.81
48	go to * restaurant	24.37	0.53	0.85
49	want to * my	24.37	0.65	0.95
50	want to * the	24.37	0.77	0.98
51	he * go to	22.94	0.25	0.71
52	the * and the	22.94	0.88	0.99
53	I think * can	22.94	0.31	0.70
54	money to * to	22.94	0.31	0.57
55	I * in the	21.50	0.73	0.96
56	I * think so	21.50	0.07	0.00
57	I * to a	21.50	0.20	0.78
58	I * to be	21.50	0.13	0.35
59	the * is not	21.50	0.67	0.95
60	I can * with	21.50	0.60	0.92
61	I think * should	21.50	0.47	0.89
62	I want * to	21.50	0.20	0.44
63	I went * the	21.50	0.07	0.00
64	it is * for	21.50	0.53	0.93
65	but * don't have	20.07	0.21	0.82
66	I * agree with	20.07	0.43	0.69
67	I * I can	20.07	0.29	0.65
68	think * is very	20.07	0.50	0.86
69	and he * to	20.07	0.71	0.97
70	I think * I	20.07	0.50	0.89
71	I want * be	20.07	0.07	0.00
72	there are * people	20.07	0.29	0.81
73	I * a lot	18.64	0.54	0.87
74	I * have a	18.64	0.31	0.74
75	I * to study	18.64	0.39	0.86
76	I * with my	18.64	0.62	0.84
77	so * have to	18.64	0.15	1.00
78	the * and I	18.64	0.92	0.99
79	there * a lot	18.64	0.23	0.78
80	go to * and	18.64	0.62	0.94
81	I don't * a	18.64	0.39	0.73
82	I don't * I	18.64	0.31	0.87
83	I don't * the	18.64	0.69	0.93
84	it is * good	18.64	0.23	0.98
85	to the * with	18.64	0.31	0.77
86	I * with it	17.20	0.17	0.92
87	I * with this	17.20	0.17	0.81
88	the * and then	17.20	0.67	0.95
89	to * with my	17.20	1.00	1.00
90	agree with * opinion	17.20	0.17	0.65
91	and I * the	17.20	0.67	0.87
92	have to * a	17.20	0.50	0.93
93	I will * to	17.20	0.42	0.84
94	in the * is	17.20	0.50	0.87
95	is a * of	17.20	0.42	0.82
96	lot of * and	17.20	0.67	0.92
97	want to * a	17.20	0.50	0.86
98	the * or the	15.77	1.00	1.00
99	have enough * to	15.77	0.27	0.55
100	I don't * with	15.77	0.27	0.55

Appendix 7. The Top 100 P-frames in Oral Production (Low-intermediate)

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
1	I think * is	127.91	0.21	0.73
2	a few * ago	126.42	0.03	0.16
3	the * of the	62.84	0.69	0.95
4	I * it is	44.13	0.07	0.23
5	so I * to	44.13	0.20	0.73
6	to the * and	43.39	0.43	0.88
7	in the * and	42.64	0.42	0.87
8	I * like to	41.14	0.11	0.52
9	I * to the	37.40	0.18	0.57
10	I think * can	37.40	0.22	0.59
11	and I * to	35.91	0.29	0.83
12	to * in the	34.41	0.46	0.86
13	I would * to	33.66	0.09	0.33
14	I * go to	32.91	0.50	0.95
15	my * and I	32.91	0.59	0.92
16	to * to the	32.17	0.30	0.53
17	with my * and	32.17	0.51	0.90
18	I * to go	31.42	0.24	0.74
19	when I * to	29.92	0.33	0.85
20	I * to do	29.17	0.21	0.68
21	I * want to	29.17	0.23	0.68
22	the * in the	29.17	0.82	0.98
23	I don't * to	29.17	0.18	0.77
24	I * I can	27.68	0.16	0.39
25	there * a lot	27.68	0.11	0.57
26	I am * to	27.68	0.57	0.89
27	I can * the	27.68	0.62	0.95
28	the * is very	26.93	0.81	0.98
29	want to * to	26.18	0.20	0.62
30	the * and I	25.43	0.82	0.97
31	I think * not	25.43	0.29	0.54
32	a * of people	24.69	0.06	0.33
33	I * to talk	24.69	0.12	0.59
34	the * and the	24.69	0.79	0.96
35	and he * to	24.69	0.49	0.90
36	to go * the	24.69	0.12	0.29
37	I don't * the	23.94	0.34	0.82
38	I went * the	23.94	0.09	0.25
39	is very * and	23.94	0.81	0.98
40	if * want to	23.19	0.19	0.78
41	and the * is	23.19	0.87	0.99
42	because I * to	23.19	0.36	0.82
43	go to * park	23.19	0.13	0.58
44	I can * my	23.19	0.71	0.96
45	to the * with	23.19	0.26	0.72
46	when I * a	23.19	0.16	0.41
47	want to * my	22.44	0.53	0.93
48	so * want to	21.69	0.10	0.65
49	are a * of	20.95	0.07	0.37
50	I don't * so	20.95	0.21	0.42
51	I think * the	20.95	0.54	0.94
52	it is * to	20.95	0.71	0.94
53	want to * the	20.95	0.68	0.97
54	he * to the	20.20	0.41	0.84
55	are * lot of	19.45	0.04	0.00

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
56	I * it's not	19.45	0.12	0.30
57	I * to work	19.45	0.27	0.77
58	when * was a	19.45	0.04	0.00
59	I think * a	19.45	0.54	0.78
60	there are * lot	19.45	0.04	0.00
61	I * in the	18.70	0.48	0.89
62	it is * good	18.70	0.28	0.78
63	my * and my	17.95	0.75	0.97
64	so * have to	17.95	0.21	0.62
65	the * with my	17.95	0.38	0.88
66	I will * to	17.95	0.50	0.87
67	is * good for	17.21	0.22	0.68
68	the * and he	17.21	0.70	0.95
69	want * go to	17.21	0.04	0.00
70	because the * is	17.21	0.78	0.96
71	have a * of	17.21	0.22	0.44
72	I think * should	17.21	0.44	0.92
73	it is * for	17.21	0.65	0.96
74	the * is not	16.46	0.86	0.99
75	think * is a	16.46	0.36	0.83
76	but I * to	16.46	0.41	0.79
77	is very * for	16.46	0.32	0.82
78	I * you to	15.71	0.10	0.45
79	very * for me	15.71	0.33	0.81
80	I have * to	15.71	0.86	0.96
81	I just * to	15.71	0.33	0.72
82	I will * my	15.71	0.52	0.89
83	want to * a	15.71	0.52	0.88
84	and * want to	14.96	0.25	0.61
85	have * lot of	14.96	0.10	0.29
86	he * to go	14.21	0.16	0.99
87	I * like the	14.21	0.32	0.67
88	the * that I	14.21	0.90	0.98
89	think * is not	14.21	0.37	0.88
90	to * with the	14.21	0.63	0.89
91	I can * to	14.21	0.63	0.94
92	I want * go	14.21	0.05	0.00
93	in the * with	14.21	0.47	0.82
94	is not * for	14.21	0.37	0.61
95	I * to have	13.47	0.28	0.60
96	I * try to	13.47	0.33	0.84
97	I * with my	13.47	0.56	0.94
98	I * with that	13.47	0.11	0.96
99	if I * to	13.47	0.39	0.84
100	is a * of	13.47	0.39	0.63

Appendix 8. The Top 100 P-frames in Oral Production (High-intermediate)

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
1	a few * ago	110.78	0.02	0.15
2	I think * is	90.91	0.13	0.67
3	the * of the	61.91	0.56	0.94
4	I * like to	52.78	0.07	0.43
5	I would * to	46.27	0.06	0.32
6	to the * and	41.05	0.38	0.82
7	I * want to	39.10	0.12	0.63
8	to * in the	39.10	0.28	0.78

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
9	in the * and	39.10	0.43	0.89
10	I * it is	37.14	0.06	0.15
11	to * to the	35.19	0.15	0.39
12	and I * to	34.86	0.21	0.79
13	I don't * to	34.21	0.09	0.65
14	to go * the	32.58	0.12	0.33
15	my * and I	31.93	0.40	0.78
16	I * to the	31.61	0.17	0.60
17	the * and the	29.98	0.77	0.96
18	with my * and	29.65	0.31	0.77
19	so I * to	27.70	0.19	0.74
20	I think * can	27.04	0.13	0.58
21	I can * my	25.42	0.53	0.90
22	because I * to	25.09	0.20	0.80
23	I think * should	25.09	0.21	0.77
24	when I * to	25.09	0.27	0.81
25	there * a lot	23.13	0.06	0.71
26	I * to go	22.81	0.23	0.81
27	I * to do	22.48	0.22	0.68
28	when I * a	22.16	0.15	0.40
29	I am * to	21.83	0.36	0.74
30	if * want to	21.51	0.14	0.67
31	the * and then	21.51	0.62	0.94
32	the * in the	21.51	0.71	0.98
33	I will * to	21.51	0.29	0.74
34	I * go to	20.53	0.40	0.91
35	it is * to	20.53	0.43	0.88
36	my * and my	19.88	0.71	0.96
37	have to * the	19.88	0.67	0.95
38	I think * the	19.88	0.33	0.85
39	have a * of	18.90	0.12	0.27
40	at the * time	18.25	0.05	0.22
41	I really * to	18.25	0.25	0.72
42	when * was a	17.92	0.04	0.13
43	it is * for	17.92	0.46	0.92
44	the * is not	17.27	0.62	0.94
45	the * is very	17.27	0.70	0.97
46	is the * important	17.27	0.04	0.14
47	have * lot of	16.94	0.02	0.00
48	I can * the	16.94	0.64	0.96
49	I just * to	16.94	0.31	0.77
50	I * going to	16.62	0.14	0.65
51	I * in the	16.62	0.57	0.92
52	the * that I	16.62	0.71	0.96
53	and the * is	16.62	0.77	0.97
54	want to * my	16.62	0.53	0.82
55	want to * to	16.62	0.28	0.69
56	when I * in	16.62	0.24	0.52
57	he * to the	15.97	0.22	0.71
58	the * and he	15.97	0.53	0.96
59	and he * to	15.97	0.27	0.89
60	I think * I	15.97	0.29	0.78
61	I went * the	15.97	0.04	0.14
62	is a * of	15.64	0.25	0.55
63	a * of people	15.31	0.04	0.15
64	so * have to	15.31	0.09	0.84
65	I don't * so	15.31	0.13	0.40

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
66	if you * to	15.31	0.19	0.56
67	is * for me	14.99	0.50	0.91
68	I can * it	14.99	0.57	0.86
69	I think * a	14.99	0.24	0.69
70	the * and I	14.66	0.80	0.97
71	want to * the	14.66	0.64	0.93
72	I * that I	14.34	0.36	0.82
73	I * try to	14.34	0.25	0.63
74	but I * to	14.34	0.36	0.81
75	want to * a	14.34	0.43	0.85
76	and * want to	14.01	0.12	0.45
77	I don't * the	14.01	0.30	0.68
78	I * a lot	13.69	0.60	0.90
79	and I * like	13.69	0.24	0.84
80	have to * a	13.69	0.60	0.94
81	is very * and	13.69	0.71	0.95
82	for me * I	13.36	0.34	0.77
83	like to * in	13.36	0.29	0.73
84	I * think so	13.03	0.05	0.17
85	it's * for me	13.03	0.50	0.92
86	think * is the	13.03	0.28	0.81
87	of the * and	13.03	0.80	0.96
88	and * is a	12.71	0.21	0.62
89	because * is a	12.71	0.31	0.76
90	there * so many	12.71	0.10	0.64
91	because the * is	12.71	0.77	0.97
92	don't know * to	12.71	0.10	0.42
93	is not * for	12.71	0.26	0.61
94	he * go to	12.38	0.21	0.69
95	I * have to	12.38	0.29	0.83
96	that * have to	12.38	0.18	0.75
97	because I * that	12.38	0.24	0.56
98	because I * the	12.38	0.34	0.82
99	is very * to	12.38	0.58	0.91
100	thank you * much	12.38	0.05	0.97

Appendix 9. The Top 100 P-frames in Oral Production (Advanced)

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
1	I think * is	132.74	0.14	0.68
2	a few * ago	104.25	0.03	0.23
3	the * of the	58.92	0.58	0.90
4	I * it is	55.68	0.05	0.20
5	in the * and	55.04	0.53	0.87
6	to the * and	47.91	0.37	0.84
7	I * like to	42.74	0.09	0.62
8	I don't * to	42.09	0.14	0.73
9	to * to the	34.97	0.20	0.51
10	so I * to	34.97	0.22	0.73
11	because I * to	32.38	0.24	0.85
12	I * want to	31.73	0.16	0.68
13	to * in the	31.73	0.35	0.82
14	I would * to	31.73	0.10	0.37
15	I * to the	30.43	0.13	0.64
16	the * and the	30.43	0.81	0.97
17	I think * can	30.43	0.21	0.70
18	a * of people	29.14	0.09	0.23

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
19	there * a lot	29.14	0.11	0.70
20	and I * to	27.84	0.33	0.83
21	I * to do	27.20	0.29	0.85
22	I can * the	27.20	0.81	0.98
23	I think * the	26.55	0.49	0.90
24	to go * the	26.55	0.10	0.25
25	he * to the	25.25	0.18	0.71
26	I * go to	24.61	0.47	0.95
27	the * and then	24.61	0.61	0.94
28	I think * have	24.61	0.29	0.80
29	and * have to	23.96	0.19	0.77
30	think * is the	23.96	0.27	0.76
31	don't know * to	23.96	0.05	0.41
32	have a * of	23.96	0.08	0.23
33	I * to say	23.31	0.14	0.67
34	my * and I	23.31	0.53	0.81
35	I think * a	23.31	0.36	0.75
36	I think * should	22.66	0.34	0.82
37	is a * of	22.66	0.26	0.59
38	I * a lot	22.02	0.77	0.96
39	the * in the	22.02	0.85	0.98
40	when I * a	22.02	0.24	0.44
41	I * to be	21.37	0.30	0.84
42	if * want to	21.37	0.15	0.69
43	and the * is	21.37	0.82	0.98
44	I * think that	20.72	0.25	0.66
45	the * and I	20.72	0.78	0.96
46	I can * my	20.72	0.59	0.93
47	I * to go	20.07	0.26	0.79
48	I am * to	20.07	0.42	0.80
49	I think * I	20.07	0.32	0.83
50	if you * to	20.07	0.29	0.58
51	is the * important	20.07	0.07	0.35
52	when I * to	20.07	0.23	0.65
53	I think * will	19.43	0.23	0.75
54	I will * to	19.43	0.47	0.85
55	lot of * and	19.43	0.87	0.98
56	it is * to	18.78	0.62	0.91
57	I * know how	18.13	0.11	0.37
58	the * is very	18.13	0.75	0.98
59	I don't * that	18.13	0.25	0.58
60	to the * with	18.13	0.32	0.80
61	want to * a	18.13	0.43	0.86
62	when I * in	18.13	0.32	0.70
63	with my * and	18.13	0.50	0.90
64	I * have a	17.48	0.33	0.87
65	and I * the	17.48	0.44	0.80
66	have the * to	17.48	0.44	0.93
67	I can * more	17.48	0.56	0.90
68	go to * beach	16.84	0.15	0.35
69	at * same time	16.19	0.04	0.00
70	can * a lot	16.19	0.64	0.94
71	I * this is	16.19	0.12	0.30
72	and he * to	16.19	0.40	0.91
73	have to * the	16.19	0.76	0.97
74	he went * the	16.19	0.12	0.40
75	I can * with	16.19	0.36	0.85

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
76	I don't * so	16.19	0.08	0.40
77	so that * can	16.19	0.24	0.76
78	the * and he	15.54	0.58	0.93
79	to * with my	15.54	0.88	0.99
80	I don't * the	15.54	0.29	0.82
81	it is * for	15.54	0.67	0.96
82	I * going to	14.89	0.13	0.76
83	I * think so	14.89	0.04	0.00
84	I * you to	14.89	0.17	0.38
85	is * lot of	14.89	0.04	0.00
86	have enough * to	14.89	0.13	0.67
87	if I * to	14.89	0.30	0.79
88	lot of * to	14.89	0.65	0.90
89	on the * and	14.89	0.87	0.99
90	want to * to	14.89	0.44	0.75
91	and * is a	14.25	0.27	0.67
92	I * have to	14.25	0.41	0.86
93	so * have to	14.25	0.23	0.64
94	and there * a	14.25	0.18	0.71
95	because I * the	14.25	0.50	0.88
96	I think * like	14.25	0.23	0.58
97	I want * to	14.25	0.14	0.34
98	I went * the	14.25	0.05	0.00
99	if I * a	14.25	0.36	0.81
100	there is * lot	14.25	0.05	0.00

Appendix 10. The Top 100 P-frames in Oral Production (Native English Speakers)

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
1	a few * ago	75.94	0.05	0.17
2	I am * to	52.57	0.26	0.79
3	I would * to	42.84	0.23	0.73
4	I * like to	40.89	0.38	0.67
5	the * that I	35.05	0.94	0.99
6	to the * and	35.05	0.67	0.94
7	I * going to	33.10	0.29	0.54
8	I * to go	31.15	0.56	0.92
9	the * and the	31.15	0.94	0.99
10	the * of the	31.15	1.00	1.00
11	I think * should	27.26	0.50	0.90
12	to go * the	27.26	0.36	0.62
13	and I * to	25.31	0.77	0.98
14	in the * and	25.31	0.92	0.99
15	you are * to	23.37	0.58	0.92
16	it * on the	21.42	0.18	0.44
17	thank * very much	21.42	0.09	0.00
18	I think * is	21.42	0.73	0.95
19	so I * to	21.42	0.73	0.97
20	go * the beach	19.47	0.20	0.47
21	I think * a	19.47	0.40	0.79
22	they are * to	19.47	0.70	0.90
23	to be * to	19.47	0.40	0.68
24	at * same time	17.52	0.11	0.00
25	he * to the	17.52	0.33	0.77
26	I * it was	17.52	0.44	0.83
27	he was * to	17.52	0.33	0.77
28	should be * to	17.52	0.33	0.62

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
29	so I * that	17.52	0.44	0.83
30	a * of people	15.58	0.25	0.54
31	a * of the	15.58	0.38	0.67
32	I * a lot	15.58	0.75	0.97
33	I * have a	15.58	0.38	0.99
34	I * it's a	15.58	0.38	0.67
35	I * to do	15.58	0.50	0.88
36	I * want to	15.58	0.63	0.93
37	my * and I	15.58	0.38	0.67
38	the * that you	15.58	1.00	1.00
39	there * a lot	15.58	0.38	0.99
40	think * would be	15.58	0.50	0.77
41	have a * of	15.58	0.50	0.77
42	I think * would	15.58	0.63	0.86
43	if I * to	15.58	0.50	0.95
44	lot of * and	15.58	1.00	1.00
45	so I * it's	15.58	0.25	0.54
46	they are * in	15.58	0.88	0.98
47	think it's * to	15.58	0.75	0.97
48	and * to the	13.63	0.71	0.92
49	as * as they	13.63	0.43	0.73
50	so * think it's	13.63	0.14	0.00
51	the * and I	13.63	0.86	0.98
52	the * that we	13.63	0.86	0.98
53	to * able to	13.63	0.14	0.00
54	to * in the	13.63	1.00	1.00
55	was * little bit	13.63	0.14	0.00
56	went * to the	13.63	0.29	0.99
57	were * in the	13.63	0.29	0.59
58	a lot * the	13.63	0.29	0.59
59	and he * a	13.63	0.86	0.98
60	and he * to	13.63	0.71	0.92
61	and I * that	13.63	0.71	0.92
62	at the * and	13.63	0.71	0.96
63	I am * a	13.63	0.43	0.73
64	I just * to	13.63	0.43	0.87
65	I think * important	13.63	0.43	0.91
66	I was * in	13.63	0.57	0.83
67	if the * is	13.63	1.00	1.00
68	if you * to	13.63	0.43	0.87
69	so I * it	13.63	0.57	0.92
70	to get * and	13.63	0.86	0.98
71	was a * bit	13.63	0.14	0.00
72	when I * a	13.63	0.29	0.59
73	a * of a	11.68	0.50	0.79
74	and * kind of	11.68	1.00	1.00
75	I * a little	11.68	0.67	0.90
76	I * as a	11.68	0.50	1.00
77	I * I have	11.68	0.50	0.92
78	I * to a	11.68	0.33	1.00
79	or * like that	11.68	0.33	0.65
80	the * and he	11.68	0.83	0.97
81	think * should be	11.68	0.67	0.90
82	you * have to	11.68	0.33	0.65
83	you * me to	11.68	0.50	0.92
84	a good * for	11.68	0.67	0.96
85	and I * like	11.68	0.67	0.90

Rank	P-frame	Normalized Frequency	VPR	Normalized Entropy
86	and the * was	11.68	1.00	1.00
87	because I * a	11.68	0.83	0.97
88	have to * the	11.68	0.67	0.90
89	I am * with	11.68	1.00	1.00
90	I think * are	11.68	0.83	0.97
91	it's very * to	11.68	0.67	0.96
92	of the * of	11.68	1.00	1.00
93	so I * a	11.68	0.83	0.97
94	to be * at	11.68	1.00	1.00
95	to go * a	11.68	0.50	0.92
96	would be * to	11.68	1.00	1.00
97	you don't * to	11.68	0.33	0.65
98	a * at the	9.74	1.00	1.00
99	I * it would	9.74	0.40	0.72
100	I * to get	9.74	0.60	0.87

THE AUTHORS

Yoon Namkung is a PhD student in the Department of Applied Linguistics and ESL at Georgia State University. Her primary research interests include the role of technology in second language development, task-based language teaching (TBLT), phraseology, and second language pragmatics.

Ute Römer is a Professor in the Department of Applied Linguistics and ESL at Georgia State University. Her primary research areas include corpus linguistics, phraseology, usage-based second language acquisition, academic discourse analysis, and the application of corpora in language learning and teaching. She serves on the editorial boards of a number of academic journals (including the *International Journal of Corpus Linguistics*, and *Corpora*) and is General Editor of the book series *Studies in Corpus Linguistics* (John Benjamins).

THE AUTHORS' ADDRESSES

First and Corresponding Author
Yoon Namkung
PhD Student
Department of Applied Linguistics & ESL
Georgia State University
25 Park Place NE, Atlanta, GA 30303, USA
E-mail: ynamkung1@gsu.edu

Co-author
Ute Römer
Professor
Department of Applied Linguistics & ESL
Georgia State University
25 Park Place NE, Suite 1500, Atlanta, GA 30303, USA
E-mail: uroemer@gsu.edu

Received: 15 October 2023
Received in Revised Form: 16 November 2023
Accepted: 8 December 2023