



# Corpus-study: Understanding Second Language Lexical Acquisition through Word Frequency Effect

Cicilia Deandra Maya Putri<sup>1</sup>

<sup>1</sup> Universitas Negeri Surabaya, Indonesia  
ciciliaputri@unesa.ac.id

**Abstract.** Corpus-based approaches have been utilized to examine second language acquisition, especially in terms of how L2 learners acquire new lexical items. The present corpus study is related to word frequency research regarding lexical item issues since the corpus provides information on how frequently a lexical item is used in natural settings. The purpose of the current study is to determine how much word frequency influences second-language lexical acquisition. Comparing word frequency from an online corpus to the real language learner word use, this study hypothesizes that word frequency could explain why the more frequently a term is used, the simpler it is for L2 learners to understand, use meaningfully, and acquire. It was found that word frequency influenced second language acquisition to an extent that second language lexical acquisition could be estimated, but was not the only predictor of second language acquisition. Variation can occur because there are other language characteristics (context, perceptual level, etc.) that affect language acquisition.

**Keywords:** Corpus Study, Second Language Acquisition, Word Frequency Effect.

## 1 Introduction

Corpus study has been used to investigate how the second language is acquired as it can provide a description of the relation between interlanguage and language development phases as well as the variation of language use in accordance with the learners' proficiency levels [1]. One of the pieces of information provided by corpus is word frequency which was argued to play a facilitative role in L2 acquisition [2-3]. The frequencies of lexical items to which learners are exposed could bring further insights into learner language and how it is developed [1, 4, 5]. Additionally, according to these studies, the more times a lexical item is used meaningfully in the receptive and production areas, the more likely it is to be incorporated into the language system. These aforementioned studies compared the word frequency effect between the language learner corpus and native speaker corpus or between two or more online corpora. However, to the extent of the author's knowledge, there is still an urgency to compare the word frequency effect obtained from online corpus to real learner language use.

### 1.1 Language Corpora

In second language development, many academics have studied how second languages are acquired, why many learners achieve different levels of proficiency, or the best approach to acquiring a second language [6]. Many have also turned to corpora dan corpus techniques to investigate how a second language is acquired as it can provide a description of the relation of interlanguage and language development phases as well as the variation of language use in accordance with the learners' proficiency levels [1]. According to Reppen [7], the corpus is a "large and principled collection of naturally occurring texts (written or spoken) stored electronically." From this perspective, it could be asserted that corpus study looks at word occurrences in natural settings, in actual language use.

The corpus-based approach has been used in language acquisition which furthermore brought many benefits. Meunier [8] researched the effect of corpus study in the language learning process and found that a balanced integration between the current multi-faceted language learning and teaching strategies and corpus linguistics is suggested to cater to certain needs for language acquisition. Another corpus-aided approach is also used to help language learners in the use of certain target language structures, for instance, Zhang and Liu [9] conducted a corpus-aided strategy to teach infinitives to Chinese EFL learners. Kettelman and Marko [10] claimed that corpus-based study promotes "learner autonomy". This is due to the fact that learners are able to draw their own conclusion about language use without having to rely solely on the teacher's intuition which can sometimes be insufficient. Hence, corpus study could be counted as one of many reliable sources for language acquisition.

### 1.2 Word Frequency Effect

Corpus study is related to word frequency research regarding lexical item issues since the corpus provides information on how frequently a lexical item is used in natural settings. Word frequency is the number of occurrences of certain words in texts. Many findings suggested that word frequency affects second or foreign language acquisition [2, 11], the so-called 'frequency effects.' Furthermore, according to Gries [12], frequency effects refer to language acquisition that is dependent on collective exposure to various linguistic objects. This phenomenon relates to the common observation that the more frequently a word occurs, the more likely that word would be acquired by L2 learners [2, 12]. Following this notion, many studies investigated word frequency's impacts on second language acquisition.

Crossley et al. [5] conducted a longitudinal study to investigate the effects of absolute word frequency in native speaker input on L2 learners in relation to L2 lexical production. They used naturalistic spoken data generated by native speakers and L2 learners to observe the frequency values for certain words. The 50 most common words shared between these variables were examined and compared to the CELEX database and the British National Corpus (BNC 2007). They found that even though the frequency effect could not interpret lexical acquisition for L2 learners, the word frequency of the native speaker input had little correlation with L2 output. This finding is in accordance with the perspective that the more learners are exposed to

certain lexical items, the easier it is for them to retain and use them meaningfully, even though this process cannot explain the order of L2 learners' lexical acquisition.

Diessel [13] researched the effects of frequency on language acquisition, language use, and diachronic change. Through summarizing and interpreting the key findings in previous research which generally compared language corpora, such as CHILDES, and Brown Corpus, to learner language, he argued that frequency reinforces the representation of linguistic expression and activation in language use. Moreover, he found that frequency plays an important role in language acquisition, though it was not the sole factor.

### 1.3 The Present Study

Crossley et al. [5] pointed out that even though word frequency could not interpret second language acquisition, it still influenced the output of L2 learners. Motivated by this finding and Diessel's [13] arguments on how frequency plays a key role in language acquisition, this present study aims to investigate the extent to which word frequency affects second-language lexical acquisition. Comparing word frequency from an online corpus to the real language learner word use, this study predicts that word frequency could explain that the more frequently a word is used, the easier it is for L2 learners to understand, use meaningfully, and acquire. However, word frequency would not be the sole predictor of second-language lexical acquisition.

## 2 Methods

Five participants were selected in this study whose L1 was not English. They came from different nationalities: Italian, Indonesian, Portuguese, German, and Dutch. They were master students of a university in the Netherlands who had approximately similar English proficiency levels, as indicated by their IELTS Scores ranging from 6.5 to 7. They were selected because of their different L1 background so that the author could see a more general result, instead of being specific to just a certain L1 background.

Word lists were prepared beforehand: 1. Stay, 2. Undergo, 3. Attack, 4. Mourn, 5. Perpetrate. These words were chosen because some of them are high-frequency words that are commonly used in conversation, such as *stay* and *attack*. The rest are low-frequency words that are relatively uncommon to be used in daily conversation, such as *mourn* and *perpetrate*. Using low and high-frequency words for this research would then prove whether high-frequency words are acquired first and low-frequency words are acquired later.

Then, to obtain the frequency numbers of these words, the study used the Corpus of Contemporary American English (COCA) [14] because it is freely accessible and it contains more than 560 million words in 220,225 texts compiled from 1990 until 2019. In addition, the corpus is evenly divided between five genres of spoken, fiction, popular magazines, newspaper, and academic journals.

To find the effect between word frequency and actual L2 learner use, this study compared the frequency numbers of a word list to the word order that was used first by L2 learners. Interviews were conducted with the five participants to find their

order of understanding and using the words in the word list. The result is captured in Table 1. Afterward, each word was typed into the COCA to look at the frequency from all five genres. The frequency data were recorded in Table 2. These data were then compared and analyzed further.

### 3 Results and Discussion

The result of the interviews can be seen in Table 1 which shows the order of lexical acquisition depending on each person. From Table 1, all participants acquired first the word 'stay', followed by 'attack', and 'undergo'. The words 'mourn' and 'perpetrate' came in last or second to last in the order of participants' lexical acquisition.

**Table 1.** Order of lexical acquisition

Participants	1 <sup>st</sup> word	2 <sup>nd</sup> word	3 <sup>rd</sup> word	4 <sup>th</sup> word	5 <sup>th</sup> word
A	stay	attack	undergo	mourn	perpetrate
B	stay	attack	undergo	mourn	perpetrate
C	stay	attack	undergo	perpetrate	mourn
D	stay	attack	undergo	mourn	perpetrate
E	stay	attack	undergo	perpetrate	mourn

Table 2 displays the number of each word occurrence (frequency) derived from COCA. Those frequencies were acquired from all five genres of spoken, fiction, popular magazines, newspaper, and academic journals. From these findings, it could already be seen that the word 'stay' appeared the most, followed by 'attack', which had tens of thousands occurrences. Their frequency numbers had very big gaps with the rest of the word list. The word 'undergo' came with way lower frequency than 'attack' and followed by 'mourn'. This left to the fact that 'perpetrate' had the lowest frequency.

The present study investigated the extent to which word frequency affects second-language lexical acquisition. Comparing word frequency from an online corpus to the real language learner lexical use, this study predicted that word frequency could explain that the more frequently a word is used, the easier it is for L2 learners to understand, use meaningfully, and acquired. However, it is not the sole predictor of second-language lexical acquisition.

With regards to the frequency effect, this study found that prediction on second language lexical acquisition was made possible by looking at the number of word occurrences from an online corpus (COCA). From the frequency of occurrences, the word 'stay' was predicted to be acquired first by L2 learners, followed by 'attack', 'undergo'. The words 'mourn', and 'perpetrate' were in interchangeable order as the last or second to last word to be acquired by the participants. Even though, there was a tendency that almost all participants had the same order of lexical acquisition as predicted by word frequencies, there were some variabilities in the order. Some parts

of this finding were consistent with previous research that the more frequently a word occurs, the more likely that word would be acquired by L2 learners [2, 12]. Diessel's [13] argument that frequency occurrence affected the processes of language acquisition was also proven in this present study. This might be because the occurrences of words lead to the activation of those words, thus they could be retained and used meaningfully in the production and receptive aspects. Thus, acquiring these words might be possible and easier.

However, looking at these findings, still, word frequency could not be the key predictor for lexical acquisition as represented by the variabilities of the two participants who had the two low-frequency words in a different order of acquisition. This was because there are influences from other linguistics aspects, such as saliency, recency, and concreteness to take into consideration. This result was consistent with the findings of Crossley et al. [5] where the absolute word frequency had little correlation with the learner's language and it solely cannot interpret lexical acquisition.

#### **4 Conclusion**

Overall, this study found that word frequency affects second-language acquisition to the extent that it could estimate second language lexical acquisition but not the sole predictor in it. Variabilities could happen because there are other linguistic properties influencing language acquisition (i.e., context, perceptual salience, etc). This study, however, is subject to several limitations. This study only used one source of corpus whereas incorporating other corpora might generate different findings. Due to time and space constraints, the study used a small sample. To be able to draw a solid and generalizable conclusion, it is suggested to involve more participants with the same or different L1 backgrounds.

#### **Authors' Contributions**

The author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

#### **Acknowledgments**

The author would like to express her heartfelt gratitude to my supervisor, Prof. Wander Lowie, for his unwavering support and guidance throughout the research process. His expertise, insights, and encouragement were invaluable in helping us to complete this work.

The author would also like to thank her colleagues at the English Literature Study Program, Faculty of Languages and Arts, Universitas Negeri Surabaya, for their helpful feedback and support.

Finally, the author would like to thank her family and friends for their love and support throughout the research process. Without their encouragement and support, the author would not have been able to complete this research.

## References

1. D. Gablasova, V. Brezina, & T. McEnery, "Exploring Learner Language Through Corpora: Comparing and Interpreting Corpus Frequency Information," *Language Learning*, 67(6), 130–154 (2017).
2. N. Ellis, "Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition," *Studies in Second Language Acquisition* 24, 143–188 (2002).
3. B. MacWhinney, "Second language acquisition and the Completion Model. In A. M. B. de Groot & J. F. Kroll (Eds.)," in *Tutorials in Bilingualism: Psycholinguistic Perspectives*. Mahwah, NJ: Lawrence Erlbaum, 113–142 (1997).
4. P. Rayson & R. Garside, "Comparing Corpora using Frequency Profiling," in *WCC '00 Proceedings of the workshop on Comparing corpora*, 1–6 (2010).
5. S. Crossley, T. Salsbury, A. Titak, & D. McNamara, "Frequency effects and second language lexical acquisition: Word types, word tokens, and word production," *International Journal of Corpus Linguistics*, 19(3), 301–332 (2014).
6. S. M. Gass & A. Mackey, "Input, Interaction, and Output in Second Language Acquisition," in *B. VanPatten & J. Williams (Eds.) Theories in Second Language Acquisition*. New York: Routledge (2014).
7. R. Reppen, *Using corpora in the language classroom*. Cambridge: Cambridge University Press (2010).
8. F. Meunier, "Corpus linguistics and second/foreign language learning: exploring multiple paths," *RBLA, Belo Horizonte*, 11(2), 459–477 (2011).
9. Y. Zhang & L. Liu, "A corpus-aided approach in EFL instruction: A case study of Chinese EFL learners' use of the infinitive," *English Language Teaching*, 7(7), 152–158 (2014).
10. B. Kettemann & G. Marko, *Teaching and Learning by Doing Corpus Analysis*. Amsterdam and New York: Rodopi (2002).
11. M. Koprowski, "Investigating the usefulness of lexical phrases in contemporary coursebooks," *ELT Journal*, 59(4), 322–332 (2005).
12. S. T. Gries, "Dispersions and adjusted frequencies in corpora," *International Journal of Corpus Linguistics*, 13(4), 403–437 (2008).
13. H. Diessel, "Frequency effects in language acquisition, language use, and diachronic change," *New Ideas in Psychology*, 25(2), 108–127, (2007).
14. M. Davies, "The Corpus of Contemporary American English (COCA) as the First Reliable Monitor Corpus of English," *Literary and Linguistic Computing*, Vol. 25, No. 4 (2010). doi:10.1093/lc/fqq018

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

