Linggle 2.0: a collocation retrieval system with quality example sentences

Shu-Li Lai[1], Jason Chang[2], Kuan-Lin Lee[3], and Wei-Chung Huang[4]

Abstract. Linggle is a pattern-based referencing tool that assists in collocation learning. In this ongoing project, we aimed to improve its performance further. First, many of the example sentences are long and difficult for students to understand, so we used a machine learning method and trained a classifier to help select dictionary-like example sentences. Second, we created a database of 60,270,000 sentences from 4C, S2ORC, and VOA Learning English. We also included Google books for real-time supplements. Then, we applied the classifier to select good example sentences from the database for display. We also limited the number of example sentences displayed for search results to improve users' experiences. Two classes of English as a Foreign Language (EFL) college students (N=51) were invited to use the enhanced tool and filled out a questionnaire. The results showed that the students were positive about Linggle's new interface and the quality of the example sentences. We expect that more EFL learners will benefit from the tool.

Keywords: collocation tool, example sentences, machine learning, interface.

1. Introduction

For EFL writers, dictionaries are essential tools that help transform learners' ideas into language. However, it is not always easy to find information on collocation and lexical grammar in a dictionary. In recent years, corpus tools have gained attention. Research has shown that corpus tools can supplement dictionaries to provide information on collocation and lexical grammar (Lai & Chen, 2015). Learners

1. National Taipei University of Business, Taipei, Taiwan; shulilai@gmail.com; https://orcid.org/0000-0002-9976-9279
2. National Tsing Hua University, Hsinchu, Taiwan; jschang@cs.nthu.edu.tw; https://orcid.org/0000-0002-8227-7382
3. National Tsing Hua University, Hsinchu, Taiwan; simon@nlplab.cc; https://orcid.org/0000-0002-9819-6755
4. National Taipei University of Business, Taipei, Taiwan; weichung.huang927@gmail.com; https://orcid.org/0000-0002-5062-6429

How to cite this article: Lai, S.-L., Chang, J., Lee, K.-L., & Huang, W.-C. (2022). Linggle 2.0: a collocation retrieval system with quality example sentences. In B. Arnbjörnsdóttir, B. Bédi, L. Bradley, K. Friðriksdóttir, H. Garðarsdóttir, S. Thouësny, & M. J. Whelpton (Eds), *Intelligent CALL, granular systems, and learner data: short papers from EUROCALL 2022* (pp. 234-239). Research-publishing.net. https://doi.org/10.14705/rpnet.2022.61.1464

can generate rules from corpus examples and learn how words and phrases are used in context (i.e. data-driven learning, see Boulton, 2017). However, observing corpus data can be time-consuming (Yoon, 2016), and it is especially challenging for students with lower English proficiency.

The Linggle collocation retrieval system is a web-based service that generates and displays information through recurring word patterns (https://linggle.com/). After typing keywords and simple syntax commands, the system shows information on collocation explicitly. Linggle has been through several revisions since the first prototype was developed (Boisson et al., 2013). In an empirical study that we conducted on the earlier version of Linggle, students found the tool easy to use and very efficient in helping them find collocation information (Lai & Chang, 2020).

However, there were some problems and room for improvement. Many of the example sentences are long and difficult to understand in the current version. Students also complained that there were too many example sentences, which overwhelmed them. Authentic sentences are often long and more complicated. In this ongoing project, we aimed to improve the quality of the example sentences to make them easier for language learners to read. Second, we also aimed to improve the interface so users would not be overwhelmed. Finally, we recruited two classes of EFL college students to evaluate the enhanced tool.

2. Method

2.1. Procedure

The Internet has plenty of texts, so it is easy to find sentences containing certain patterns or collocations. However, some sentences may not be appropriate for language learners because they are too difficult to understand. Inspired by the GDEX model (Kilgarriff et al., 2008), we originally planned to take a rule-based approach to extract sentences from the web. For example, limit the sentence length, eliminate non-alphanumeric characters, and exclude sentences containing blacklist words. However, we found that sentences fitting the rules were not necessarily good example sentences. Some were still difficult to comprehend and contained difficult words. Our project took a different approach: we used the machine learning method and trained a classifier to help select dictionary-like examples automatically.

We used the pre-trained BERT model to extract features for the classifier. Positive data included 140,000 example sentences from the Cambridge dictionary and negative data comprised a random sample of 140,000 example sentences from Wikipedia, as those sentences are typically not as good as sentences from the learners' dictionaries for language learning purposes. The two sets of sentences were combined to create a dataset for training, developing, and testing. For training and development purposes, we randomly split the dataset into a training set (80%), a development set (10%), and a test set (10%). After training, the test set was used to evaluate the performance of the model. The average accuracy was 0.95, indicating that the model was capable of distinguishing between good (dictionary-like) and bad (not dictionary-like) sentences. Then, we applied this model to a collection of open-source corpora to create an example dataset for retrieving and presenting quality example sentences for words and phrases in our enhanced collocation retrieval system.

2.2. Corpus

We used 60,270,000 sentences from four sources to compile our dataset of example sentences: C4, S2ORC, VOA Learning English, and Google books. We sampled about 30 million sentences from the subset of 'realnewslike' in Google's C4, a colossal corpus of cleaned web crawl data, to expand coverage of the topic; 30 million sentences from S2ORC, an enormous corpus of scientific journals from the Allen Institute for AI; 270,000 sentences from VOA Learning English; and Google books for real-time supplement sentences if there were any phrases not covered by the previous three corpora.

2.3. Interface

The major modification was the restriction of example phrases for each collocation combination. Using 'to gain knowledge' as an example, the upgraded 2.0 version of the system presented five example phrases to users first. If the user clicked on 'show more', the system displayed five additional sentences, as shown in Figure 1 and Figure 2 below.

2.4. Questionnaire

We conducted a small-scale study with 51 EFL college students to collect their initial feedback. The participants' proficiency levels ranged from A2 to B2. Three class meetings were arranged to orient the students to learn the concept of corpus, collocation, and the syntax commands of Linggle. A six-point Likert scale

questionnaire was designed to understand their experiences in using the upgraded version of the system and to elicit comments.

Figure 1. Example sentences for ‘to gain knowledge’

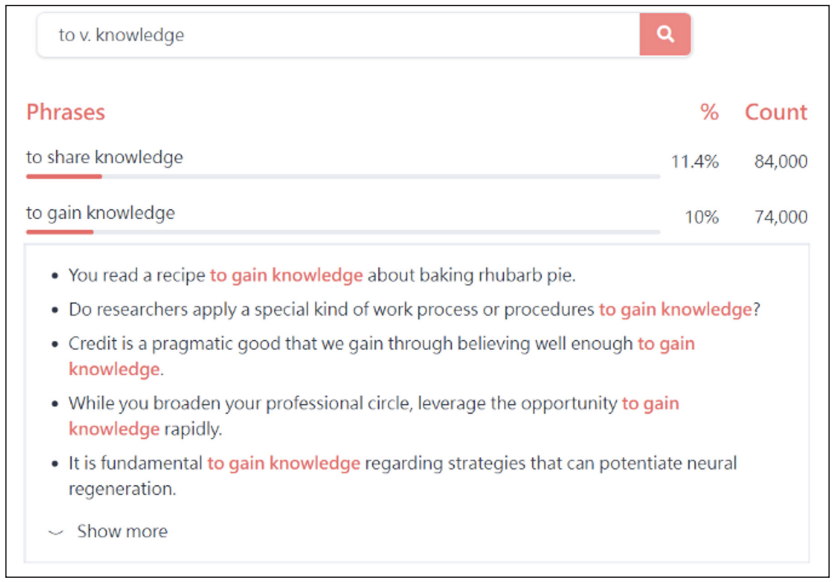
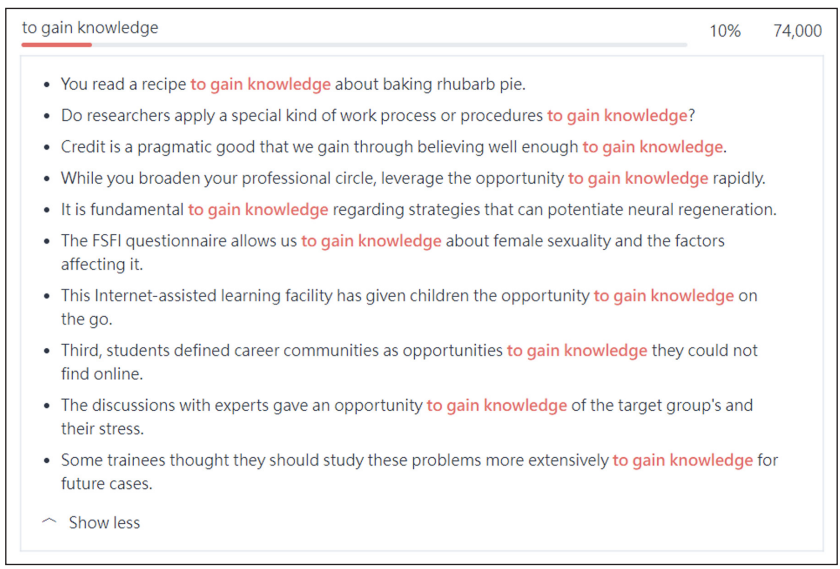


Figure 2. More example sentences for ‘to gain knowledge’



3. Results and discussion

Table 1 shows the results of the participants' initial experiences with the improved system. In general, students found Linggle highly efficient. It helped them with most of the collocation problems (*M*=5.3). Students also found Linggle easy to use (*M*=5.3) and the syntax commands are easy to learn (*M*=5.2). The findings are consistent with the previous Linggle study (Lai & Chang, 2020).

Table 1. Questionnaire results on efficiency and usability of the tool

	Item	AVG
1	Linggle helped me with most of the collocation problems.	5.3
2	Linggle quickly helped me solve collocation problems.	5.1
3	I find it easy to use Linggle to look up collocations.	5.3
4	The Linggle syntax commands are easy to learn.	5.2

Table 2 shows the participants' perceptions of the new interface and the information the system provided, including the collocation information and the example sentences for each collocation. The students were pleased with the interface, which was clear, intuitive, and easy to use. They also believed that the enhanced version offered a reasonable number of example sentences (*M*=4.5) and these sentences were not too hard (*M*=4.7). Observing corpus examples is never easy. Learners often experienced some frustration when observing corpus examples (Yoon, 2016). In this project, the example sentences were selected by the classifier we trained. The dictionary-like sentences were easier for the EFL learners to comprehend and would reduce learners' cognitive load.

Table 2. Questionnaire results on interface and richness of example sentences

	Item	AVG
1	Linggle's interface is user-friendly.	5.4
2	I like how Linggle displays its search results.	5.2
3	Linggle offers a lot of information on patterns and collocations.	5.2
4	The examples that Linggle gives are not too hard or too easy.	4.7
5	Linggle offers a reasonable number of example sentences.	4.5

4. Conclusion

This paper reported the improvements made to the collocation search engine Linggle. We used the machine learning method and trained a classifier to help select good example sentences automatically. Initial results were quite positive. We

also enhanced the interface to avoid overwhelming users. If a corpus tool is well-designed, it can be very beneficial. Our enhanced version will provide EFL learners a quick way to obtain collocation information.

5. Acknowledgements

The project was sponsored by the Ministry of Science and Technology, Taiwan (MOST 110-2637-H-141-001).

References

Boisson, J., Kao, T.-H., Wu, J.-C., Yen, T.-H., & Chang, J. S. (2013) Linggle: a web-scale linguistic search engine for words in context. In H. Schuetze, P. Fung & M. Poesio (Eds), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 139-144).

Boulton, A. (2017). Data-driven learning and language pedagogy. In S. Thorne & S. May (Eds), *Language, education and technology: encyclopedia of language and education*. Springer. https://doi.org/10.1007/978-3-319-02328-1_15-1

Kilgarriff, A., Husak, M., McAdam, K., Rundell, M., & Rychly, P. (2008). GDEX: automatically finding good dictionary examples in a corpus. In E. Bernal & J. DeCesaris (Eds), *Proceedings of the XIII EURALEX International Congress*. Universitat Pompeu Fabra.

Lai, S. L., & Chang, J. S. (2020). Toward a pattern-based referencing tool: learner interactions and perceptions. *ReCALL*, 32(3), 272-290. https://doi.org/10.1017/S0958344020000105

Lai, S. L, & Chen, H. H. (2015). Dictionaries vs concordancers: actual practice of the two different tools in EFL writing. *Computer Assisted Language Learning*, *28*(4), 341-363. https://doi.org/10.1080/09588221.2013.839567

Yoon, C. (2016) Concordancers and dictionaries as problem-solving tools for ESL academic writing. *Language Learning & Technology*, 20(1), 209-229. https://doi.org/10125/44453

Published by Research-publishing.net, a not-for-profit association
Contact: info@research-publishing.net

Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022
Edited by Birna Arnbjörnsdóttir, Branislav Bédi, Linda Bradley, Kolbrún Friðriksdóttir, Hólmfríður Garðarsdóttir, Sylvie Thouësny, and Matthew James Whelpton

Publication date: 2022/12/12

Typeset by Research-publishing.net
Cover photo by © 2022 Kristinn Ingvarsson (photo is taken inside Veröld – House of Vigdís)
Cover layout by © 2022 Raphaël Savina (raphael@savina.net)

ISBN13: 978-2-38372-015-7 (PDF, colour)

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2022.