Data-Driven Learning and Low-Level Learners

Barney Meekin

It has been claimed that data-driven learning (DDL) has many benefits for second language acquisition. However, due to the difficulty of the language included in corpora and the complexity of concordancing software, the suitability of DDL for low-level learners has been questioned. This study investigated the appropriateness of DDL through direct access to a corpus in high school lessons and compared the use of a graded and a non-graded corpus. Through parametric testing, results showed that there was no statistically significant difference between improvements made by learners accessing either a graded or a non-graded corpus. However, analysis of a post-study questionnaire suggests that a non-graded corpus was more positively perceived by learners. The study has identified some interesting areas for future, larger-scale research into DDL. Despite the challenges of DDL, directly accessing corpora can be an effective and meaningful way for learners to notice and discover elements of authentic English in use.

データ駆動型学習 (DDL) は、第二言語習得に多くの利点があるといわれているが、コーパスに含まれる言語の難しさやコンコーダンスソフトの複雑さから、初級学習者にとってDDLがふさわしいかどうかは疑問視されている。本研究では、高校の授業において、コーパスへ直接アクセスするDDLが適切かどうかを調査し、語彙制限コーパスと非制限コーパスの使用を比較した。パラメトリック検定の結果、語彙制限コーパスと非制限コーパスの間には統計的に有意な差はみられなかったが、学習後のアンケート調査では、非制限コーパスの方が学習者に肯定的に受け入れられていることが示唆された。本研究により、DDLに関する今後の大規模な研究に向けて興味深い領域がいくつか明らかになった。DDLの難しさにもかかわらず、コーパスに直接アクセスすることは、学習者が生きた英語の要素に気づき、発見するための効果的で有意義な方法であると考えられる。

Data-driven learning is "a student-centred inductive method of language learning, in which learners explore grammar and vocabulary issues using a corpus."(Hadley & Charles, 2017, p. 131). Specialised software allows learners to analyse corpus data in terms of language usage, frequency, collocations, and multi-word units both quantitatively and qualitatively (Friginal, 2018). By using concordancing functions and key words in context (KWIC), learners enquire and speculate, becoming aware of language features and reaching their own conclusions and hypotheses about language use and meaning (Friginal, 2018), which increases language awareness and learner autonomy (Boulton, 2009). However, despite research showing that DDL can be beneficial to intermediate and advanced learners, factors including the language difficulty and distractions caused by the decontextualised nature of KWIC have led to mixed results with lower level learners (Hadley & Charles, 2017). The aim of this small-scale study was to explore how DDL can be effectively used in low-level English lessons, to identify differences between using a graded and a non-graded corpus, and to generate research questions for future, larger-scale research.

Background

DDL Benefits

There are several advantages of using DDL in the language classroom. Firstly, it offers the opportunity for learners to notice features of language in use, which is a vital component of vocabulary acquisition (Nation, 2003). In Schmidt's influential Noticing Hypothesis (1990), input becomes intake after being noticed by the learner. Due to the opportunities of such conscious noticing given by DDL, it has a solid basis in second-language acquisition theory. Another major benefit of DDL is that by giving learners access to a corpus, they are provided with a much greater number of authentic texts within which they are able to view and analyse a large number of examples of language in use (Reppen, 2011). Learners can use this wealth of data to notice features of language through the examples in context (McNair, 2018) displayed by the concordancing software in a learner-centred and inductive environment (Boulton, 2011). Furthermore, by attempting to understand language features through their contexts, learners can see the connections with the words to the left and the right rather than considering lexical items individually (Friginal, 2018). Meeting words in a variety of contexts, rather than in decontextualized examples, may lead to the acquisition of "rich, transferable knowledge" (Cobb, 1997, p. 303), thus "offering the potential for deep word knowledge" (Allan, 2009, p. 24). DDL is most beneficial, with regard to the acquisition of vocabulary (Gilquin & Granger, 2010), as learners are given the opportunity to encounter words in lexical bundles with their collocations. There has been a recent realisation of how important lexical bundles are in language teaching with evidence showing that learning lexical phrases as wholes, rather than as individual words, has positive effects on language acquisition (Allan, 2016). For example, Shin and

Nation (2008) claim that having language chunks at one's disposal reduces processing time and requires less cognitive effort, therefore leading to improved fluency. They also claim that the learning of lexical bundles helps with the problem of grammatically correct utterances which include unnatural collocations and word selection. Finally, because roles in the DDL classroom are changed, the teacher becomes a facilitator (Nolan, 2018), guiding the learners and scaffolding the activities, ultimately allowing learners to make their own discoveries regarding the language and self-direct and take responsibility for their own learning (Boulton, 2011). This element of discovery present in DDL can increase enjoyment, cognitive engagement, autonomy, and empowerment, which can lead to confident learners who are more akin to "travellers," "researchers," and "detectives" (Gilquin & Granger, 2010).

Limitations

Although DDL is supported by Schmidt's Noticing Hypothesis (1990), it may be in conflict with Krashen's Input Hypothesis and Vygotsky's zone of proximal development as by accessing a large corpus made up of natural English from a variety of sources, learners may encounter language far above their current level of competency (Hadley & Charles, 2017). This may be one of the reasons why DDL is often seen as unsuitable for lower-level learners (Friginal, 2018). Because concordancing technology can be intimidating for both learners and teachers due to its complexity, DDL is a time-consuming methodology. Not only can it take weeks to train learners how to use software successfully (Friginal, 2018), as teachers of DDL lessons are required to train students in the use of concordancing software and assist with any issues which may arise (Nolan, 2018), time must also be taken to train teachers. Friginal (2018) also claims that the reading of KWIC lines may be tedious for learners and could inhibit learning. Similarly, the switching of roles could lead to issues with cultural differences. For example, learners belonging to cultures in which education is traditionally teacher-centred may struggle to self-direct their learning. Again, it may take considerable time to train such learners to reap the benefits of DDL.

Method

This study was done as an introduction to DDL in a mid-level private Japanese high school for a total of eight one-hour classes. The participants for the study were the author's first grade English class, which was made up of 31 16-year-old students. All students had at least three and a half years formal English education by the start of the study and were considered to be around level A1 or A2 on the CEFR-J scale. The class was regarded as the highest level academically of the first-grade classes. During their English lessons for the previous eight months, the participants had taken part in an extensive reading program and had many opportunities to participate in project- and task-based learning. Therefore, they were already relatively autonomous learners of English who were able to self-direct their study. In addition, all students in the class had access to their own tablet devices enabling the direct access of online corpora. The group responded well to group activities and enjoyed being involved in some kind of competition. However, because they were focused on university entrance exams, they sometimes failed to see the benefit of classes not including the discrete teaching of grammar points or grammar practice drills. This was made clear to the teacher through informal discussions.

Corpora

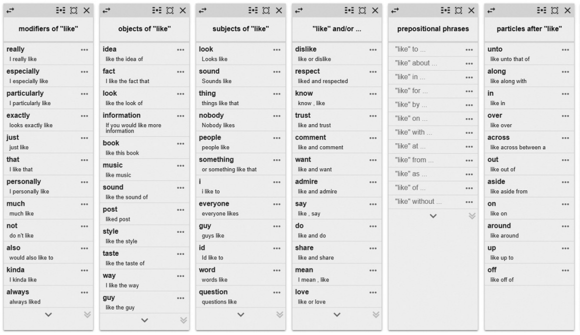
In order to mitigate some of the issues surrounding the difficulty of using concordancing software, Sketch Engine was chosen as the platform for this research due to its accessible user interface and the clear and easy-to-understand way results are displayed. Sketch Engine allows users to identify the most frequent word in a corpus with the Wordlist function, strong combinations of words with the Word Sketch (see Figures 1 & 2) function, and view KWIC through the Concordance function (Sketch Engine, 2020). Despite Sketch Engine having several open corpora which can be accessed freely, for this research the researcher and participants were given full-access accounts for the length of the study due to the researcher's affiliation with a university within the EU. Because of this, the researcher was able to create an original corpus on the Sketch Engine platform and give participants access to a non-open corpus which would not have been possible with a free account.

Participants were divided into two groups and given access to different corpora. Group 1, made of 15 participants, was given access to a corpus specially made by the author. In an attempt to alleviate the issues regarding comprehensible input, the corpus was graded to an understandable level. According to Allan (2009), the length of sentences and number of infrequent words used in native discourse are problematic for learners. Allan suggests the use of graded corpora to negate this issue because they supply a sufficient and similar range of lexical bundles to ungraded texts but are more understandable for learners (Allan, 2016). Therefore, Group

1's corpus included a range of news articles from online sources which provide graded texts such as Newsela and Breaking News English. Furthermore, two free non-fiction mid-frequency (Nation & Anthony, 2013) graded readers at the 4,000-word level were included. Primarily, these sources were chosen due to their availability. In spite of needing a very large corpus for linguistical research, many classroom practitioners prefer to use a smaller, more manageable corpus in their classes although there is disagreement about how big one should actually be (Chambers, 2007). The ranges in size quoted in Chamber's research span from 20,000 to 1 million words. The decision was made to limit the size of this corpus in order to provide fewer but more relevant examples to learners. The final total size of Group 1's corpus was 233,721 words.

Figure 1

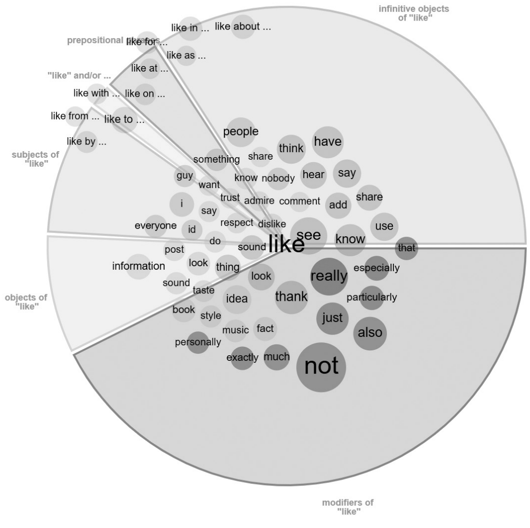
Example Output from Sketch Engine Word Sketch Function



modifiers of "like"	objects of "like"	subjects of "like"	"like" and/or ...	prepositional phrases	particles after "like"
really I really like	**idea** like the idea of	**look** Looks like	**dislike** like or dislike	"like" to ...	**unto** like unto that of
especially I especially like	**fact** I like the fact that	**sound** Sounds like	**respect** liked and respected	"like" about ...	**along** like along with
particularly I particularly like	**look** like the look of	**thing** things like that	**know** know , like	"like" in ...	**in** like in
exactly looks exactly like	**information** If you would like more information	**nobody** Nobody likes	**trust** like and trust	"like" for ...	**over** like over
just just like	**book** like this book	**people** people like	**comment** like and comment	"like" by ...	**across** like across between a
that I like that	**music** like music	**something** or something like that	**want** like and want	"like" on ...	**out** like out of
personally I personally like	**sound** like the sound of	**i** i like to	**admire** like and admire	"like" with ...	**aside** like aside from
much much like	**post** liked post	**everyone** everyone likes	**say** like , say	"like" at ...	**on** like on
not do n't like	**style** like the style	**guy** guys like	**do** like and do	"like" from ...	**around** like around
also would also like to	**taste** like the taste of	**id** Id like to	**share** like and share	"like" as ...	**up** like up to
kinda I kinda like	**way** I like the way	**word** words like	**mean** I mean , like	"like" of ...	**off** like off of
always always liked	**guy** like the guy	**question** questions like	**love** like or love	"like" without ...	

Figure 2

Sketch Engine Word Sketch Visualisation Tool



Group 2, made of 16 participants, had access to a large, non-graded corpus: SiBol: Corpus of English broadsheet newspapers 1993–2013. This corpus includes news articles from a variety of English-language newspapers over a 20-year period; it totals around 650 million words from 1.5 million articles from 14 broadsheet newspapers around the world (Sketch Engine, 2018). This corpus was chosen in an attempt to somewhat match the content of the two corpora.

Activities

Activities for the DDL lessons were developed to both try and mitigate some of the previously mentioned limitations of DDL whilst at the same time facilitating the benefits. To mitigate the difficulty of concordancing software, all activities were scaffolded with a clear progression from simple, teacher-led tasks revolving around the less complicated of Sketch Engine's functions to more complex and learner-led ones at the end of the study. The early activities, designed to introduce the concordancing software to participants, were gamified to introduce an element of competition to the lessons and all tasks were designed to be collaborative, often in groups of three to seven people, which afforded weaker students the option of staying relatively hidden. In order to facilitate the benefits of DDL, all tasks encouraged the noticing of language features and patterns necessary for language acquisition (Nation, 2013) and involved aspects of discovery learning. To create relevancy with the research and their other studies, DDL tasks were linked to the extensive reading program. For example, after making notes of words or phrases they did not understand from recent sustained-silent reading, learners used the corpora, instead of a dictionary, to check their understanding. For the final activity of the research, learners researched a word or lexical bundle of their own choosing and displayed their findings as a free-form presentation. Due to the lack of restrictions given to the type of presentation they could do, learners had the opportunity for creative thinking and output with some using their allotted time to teach mini lessons, some creating skits, and others focusing on the different, more abstract meanings of ostensibly simple words.

Before and after DDL lessons, learners were asked to complete, under exam conditions, the monolingual version of Nation and Beglar's (2007) Vocabulary Size Test. This receptive vocabulary test, which measures vocabulary knowledge, consists of 140 multiple-choice questions from the 14,000 most frequent English words. This test was chosen due to its reliability and consistency, the ease in which it can be scored and interpreted, its appropriateness for learners of all levels, its lack of ambiguity within the items (Nation & Beglar, 2007), and its replicability. Parametric tests were then conducted on the test scores to identify any patterns. In addition, an anonymous survey made up of nine items, including Likert items and multiple-choice questions (Table 1), was administered at the end of the research. Several of the items included a space in which participants could record reasons for their responses.

Table 1

Questionnaire

Did you enjoy the lessons using Sketch Engine? Why (not)?
Do you think using the corpus in class was a waste of time?
How useful was the corpus for your English learning?
Did you learn any new words?
Compared to a dictionary, how useful was the corpus? Why?
Would you rather use a dictionary or a corpus? Why?
How do you feel about the amount of data available in your corpus?
Were the examples difficult to understand?
Would you like to do lessons of this kind again? Why (not)?

Results and Discussion

Pre- and Post-Test Results

Descriptive statistics for pre- and post-tests were calculated (Table 2) and compared. To compare the scores, in order to identify if there was a statistically significant difference, paired sample t-tests were conducted for both groups. In terms of Group 1, no statistical difference was found between the pre- (*M* = 52, *SD* = 11.1) and the post-test (*M* = 53.33, *SD* = 7.97), *t*(14) = -.57, *p* = .58; therefore, the null hypothesis was not rejected. On the other hand, a significant difference was found between Group 2's pre- (*M* = 50.69, *SD* = 7.01) and post-test results (*M* = 55.25, *SD* = 10.56), *t*(15) = -2.34, *p*<.05. Therefore, the null hypothesis was rejected. Furthermore, a medium effect size (*d* = .584) was found. Results here imply that having access to a larger, non-graded corpus led to better post-test results. This is interesting as it seems to contradict previous research claiming that such a large corpus may have negative effects when compared with a smaller corpus (Chambers, 2007).

Table 2

Pre- and Post-Test Descriptive Statistics

Group	Test	*M*	*SD*	*Min*	*Max*	*n*
1	Pre	52	11.1	37	80	15
1	Post	53.33	7.97	38	71	15
2	Pre	50.69	7.01	40	65	16
2	Post	55.25	10.56	38	75	16

In order to confirm if this was the case, an analysis of covariance (ANCOVA) was conducted. ANCOVA is able to indicate whether there was any significant difference between the improvements made by Group 1 and Group 2. Results from the ANCOVA indicated that there was no significant difference between the post-test improvements of Group 1 and Group 2, *F* (1, 28) = .94, *p* = .34 while adjusting for pre-test scores. Therefore, despite the significant difference found in Group 2's paired sample t-tests, the kind of corpus learners accessed had no statistically significant effect on post-test improvement.

Questionnaire Results

Analysis of the questionnaire responses can give us further insight into the success of the research lessons. To begin with, questions related to overall enjoyment of the lessons were analysed. Responses for item one showed that 83.8% (*n* = 26) of participants enjoyed the DDL lessons. Of these 26 participants, 12 (80%) belonged to Group 1 and 14 (87.5%) to Group 2. Reasons given for enjoying the DDL included the novelty of accessing a corpus, working in a team with friends, and enjoying using their devices in class. For example, when talking about the novelty of the lessons, one participant wrote, "Because learning new methods that I don't know is really interesting [sic]" and another wrote, "it is more fun than i study only words [sic]." Responses for item nine revealed that only six participants (40%) of Group 1 would like to do DDL lessons in the future. On the other hand, 11 participants (68.8%) of Group 2 answered the same question positively. When asked for reasons why they would or would not like to try lessons like that again, some participants claimed it was a good tool to help their understanding of vocabulary.

However, others claimed they believed grammar instruction would be more valuable. Some claimed they would be willing to do it again as long as they could combine the corpus with dictionary use, and one, from Group 1, stated they thought it would be better if the corpus included a language level more similar to the level of the books they had access to in the school's extensive reading program. These results indicate that participants saw some benefit to DDL. However, they would like it to be combined with more traditional second language teaching methods. Furthermore, due to these particular learners being focused on test scores, without a clear link between DDL and grades, motivation may be negatively affected. For example, one participant responded with "i want to get score of my test and answer [sic]." Overall, from analysis of these two questions, we can see that participants in Group 2 had a more positive experience in the DDL lessons despite the higher difficulty of their corpus.

Next, questions related to the appropriateness and effectiveness of DDL were analysed. Question 2 showed that 53.33% (n = 8) of Group 1 and 81.25% (n = 13) of Group 2 felt the DDL lessons were not a waste of class time. Despite both Group 1 and Group 2 (93.33% and 81.25% respectively) indicating that they had learned new words during the DDL lessons, only 40% of Group 1 (n = 6) and 50% of Group 2 (n = 8) felt that the corpus was a useful tool for learning English. Furthermore, the majority of participants (80% for Group 1 and 68.75% of Group 2) stated they would rather use a dictionary than a corpus for vocabulary learning. One participant wrote, "I don't have enough time for using SketchEngine but I think very good for person have much time and can understand words [sic]" and another wrote, "It is also important to search and understand words quickly for learning. At this point, dictionary is better. But for writting a essay, corpus is more useful than dictionary because it shows variety of examples [sic]." The amount of time and effort taken to use the concordancing software was a common complaint. For example, a Group 2 participant wrote, "If I use SketchEngine, I need much time to understand meanings [sic]." This indicates that the participants may be unable to envision adding concordancing software study to their already busy school schedules. Finally, to assess the appropriateness of a graded or non-graded corpus, questions 7 and 8 were analysed. Understandably, 68.75% of Group 2, who had access to the ungraded corpus, felt there was too much data in their corpus. This is in contrast to just 33.33% of Group 1 who felt the same way. 73.33% and 81.25% of Group 1 and 2 respectively, claimed the examples in their corpora were too difficult. For the open-ended questions, there were similar results regarding the difficulty of the corpus examples with one Group 2 participant writing, "Because I do not know enough vocabulary to understand example sentences so I can not understand the ward I was interesting [sic]."

Questionnaire results have revealed some interesting implications. Participants enjoyed the gamified nature of the DDL lessons and access to new technology and style of lessons, but many of the limits previously described about the difficulty of language, complexity of software, and amount of data are valid concerns. However, the participants' responses imply that combining DDL with more traditional features of education such as discretely taught grammar, formal assessments and grades, or combining DDL with productive writing tasks might mitigate the difficulties. For instance, integrating corpus linguistics with traditional writing lessons would allow learners to discover not only word combinations but different types of sentence structures, which could then be linked to paraphrasing skills or output tasks whilst also meeting more traditional educational needs. Although it would seem a non-graded corpus was more successful, there is one caveat: The size of the corpus seems to have been daunting for some participants. Rather than controlling the level of the language and using only graded texts, it may be more beneficial to use native-level language, but control the amount of data.

Conclusion

The aim of this small-scale study was to explore how DDL can be used effectively with low-level learners and to identify some interesting areas for future research. Despite previous research suggesting that a non-graded corpus may have negative effects (Chambers, 2007), results from parametric testing showed no significant difference between the two groups. Furthermore, results from the questionnaire imply that participants who used the non-graded corpus found the research lessons more meaningful despite their concerns about the size of the corpus. By limiting the size of a corpus rather than the level of language, this could be mitigated. Further research is required to assess this. Namely, for low-level learners, what is more effective: a graded corpus, an ungraded corpus with a limited number of examples, or an ungraded corpus with no restrictions placed on its size?

For Japanese high school students who are focused on university entrance exams consisting of decontextualised grammar and vocabulary questions, DDL may not feel relevant, and there were questionnaire responses in this study suggesting that DDL had not satisfied the participants' educational needs. This paper suggested DDL could be integrated into writing lessons and connected with traditional academic skills such as paraphrasing, but questions still remain about how to effectively combine DDL and more traditional educational goals for many Japanese high school students. Here is another interesting area that deserves to be looked at in larger-scale research: How can DDL be effectively combined with the traditional needs of Japanese high school learners?

The current study has identified some benefits and limitations of DDL in a low-level class in line with the mixed results from previous research. Overall, the participants had positive experiences during the DDL lessons, but indicated that they were overwhelmed by the amount of data available, the difficulty of the language, and the complexity of the technology. Despite this, DDL can be a useful tool for teachers to encourage discovery, autonomy, and a deeper understanding of language even in the low-level classroom.

References

Allan, R. (2009). Can a graded reader corpus provide 'authentic'input? *ELT Journal, 63*(1), 23–32. https://doi.org/10.1093/elt/ccn011

Allan, R. (2016). Lexical bundles in graded readers: To what extent does language restriction affect lexical patterning? *System, 59*, 61–72. https://doi.org/10.1016/j.system.2016.04.005

Boulton, A. (2009). Testing the limits of data-driven learning: Language proficiency and training. *ReCALL, 21*(1), 37–54. doi: https://doi.org/10.1017/S0958344009000068

Boulton, A. (2011). Data-driven learning: the perpetual enigma. In S. Goźdź-Roszkowski (Ed.), *Explorations Across Languages and Corpora* (pp. 563–580). Peter Lang.

Chambers, A. (2007). Popularising corpus consultation by language learners and teachers. In *Corpora in the Foreign Language Classroom* (pp. 1–16). Brill Rodopi.

Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System, 25*(3), 301–315. https://doi.org/10.1016/S0346-251X(97)00024-9

Friginal, E. (2018). *Corpus Linguistics for English Teachers: Tools, Online Resources, and Classroom Activities*. Routledge.

Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language teaching. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 359–370). Routledge.

Hadley, G., & Charles, M. (2017). Enhancing extensive reading with data-driven learning. *Language Learning & Technology, 21*(3), 131–152. https://doi.org/10.3389/feduc.2019.00007

McNair, J. (2018). Using a concordance for vocabulary learning with pre-intermediate EFL students. In E. Friginal, *Corpus Linguistics for English Teachers. New tools, Online Resources, and Classroom Activities* (pp. 224–232). Routledge.

Nation, P. (2003). Materials for teaching vocabulary. In B. Tomlinson (Ed.), *Developing Materials for Language Teaching* (pp. 394–405). Bloomsbury Publishing.

Nation, P., & Anthony, L. (2013). Mid-frequency readers. *Journal of Extensive Reading, 1*, 5–16.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Nolan, M. (2018). Experiencing CL and DDL in Practice through Phraseology. In E. Friginal (Ed.), *Corpus Linguistics for English Teachers: New Tools Online Resources, and Classroom Activities* (pp. 43–45). Routledge.

Reppen, R. (2011). Using corpora in the language classroom. In B. Tomlinson (Ed.), *Materials Development in Language Teaching* (pp. 35–51).

Schmidt, R. W. (1990). The role of consciousness in second language learning[1]. *Applied Linguistics, 11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Shin, D., & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal, 62*(4), 339–348. https://doi.org/10.1093/elt/ccm091

Sketch Engine. (2018). SiBol: Corpus of English broadsheet newspapers 1993-2013. Retrieved from https://www.sketchengine.eu/sibol-corpus/

Sketch Engine. (2020). How to use a corpus to get information about words. Retrieved from https://www.sketchengine.eu/what-can-sketch-engine-do/

Barney Meekin has been teaching English since first moving to Japan in 2007. He's also taught general English in Australia, and IELTS and EAP in New Zealand. In addition to CELTA and Delta qualifications, he holds an MA in TESOL and applied linguistics from the University of Leicester. His research interests include student engagement, PBL and social emotional learning. He is currently teaching at several schools in the Kansai area.