Insights from corpus linguistics: Using keywords-in-context to teach and assess English learning

Rodrigo Arellano & Kevin Gerigk

Introduction

A common problem when teaching English is the lack of systematicity when presenting vocabulary or evaluating student language use. In some cases, the language presented in teacher examples may not be representative of either authentic input or the language utilised by learners in assessment practices. Furthermore, most ready-made teaching materials follow pre-selected topics, which may diverge from the learners' individual language requirements, resulting in a lack of applicability and motivation (Gerigk, 2022).

In this article, we propose a corpus linguistics approach to language teaching and assessment. A corpus is a large, principled collection of texts, which can be used for linguistic analysis (McEnery & Hardie, 2012). Today, corpora are available that represent different domains of language use, such as spoken and written corpora (e.g., the British National Corpus 2014; Love et al., 2017), or different genres, such as newspaper articles from the web and academic lectures or student essays (e.g., British Academic Writing Corpus). These corpora offer insight into language-in-use in its respective context. Whilst most language teaching materials are still based on the intuitions of materials writers, corpora may provide teachers with "actual evidence of [language] use" (O'Keeffe et al., 2007, p. 21). This may lead to more authentic language teaching and has the potential to increase learner and teacher autonomy (Charles, 2014).

This short article introduces three activities that operationalise corpus-based, data-driven learning and offer an opportunity for learners and teachers to discover language patterns (Timmis, 2015). The activities are particularly aimed at vocabulary acquisition and writing feedback.

Activity 1: Reported informal speech in conversational English

This activity makes use of BNClab (http://corpora.lancs.ac.uk/bnclab/), an online, freeware corpus interface hosted by Lancaster University, which allows users to explore the British National Corpus to access a total of 200 million words of English.

We have two options when reporting on something someone mentioned: 1) direct

speech (He said, 'I am going home'), or 2) indirect speech (He said he was going home). However, in naturally occurring conversations, this is not always observed.

In the following task, the phrase '(s)he was like' is introduced through an exploratory approach as an informal way of reporting speech.

Exercise 1:

- Access BNClab and enter the phrases 'he said' and 'he was like' into the search bar.
- Compare the frequencies for both spoken and written English.

Exercise 2:

- Repeat the above steps but tick the box *Age*.
- What differences can you spot between the two phrases?
- Is there a preference for one phrase over another by age group?

Exercise 3:

- Search the phrase 'he was like' and extend the context to 50 words.
- What observations can you make in terms of formality and back-shifting tenses?

Exercise 4:

- Return to the main menu and run a query for 'she was like' and 'she said'.
- For each of the phrases, navigate to *Change* and find out about their frequency between the 1990s and 2010s.
- How does that relate to the findings regarding *Age* in Exercise 2 above?

Activity 2: Semantic collocates of quasi-synonyms

This activity uses SketchEngine (https://www.sketchengine.eu/), which is a low-cost, subscription-based corpus analysis software with a web interface, as it offers access to more than 600 corpora from roughly 90 languages. There is a free trial version of SketchEngine available.

The spoken British National Corpus 2014 can be accessed via SketchEngine and has been used for the following task.

In language learning, we are often confronted with synonyms and collocations, as equivalent lexical items or words that often co-occur together, respectively (O'Keeffe

et al., 2007). However, violating their use may lead to the phrase sounding odd. This activity looks at the two quasi-synonymous verbs 'to get' and 'to become' by investigating their contextual occurrences using the Word Sketch function in SketchEngine to create various visualisations.

Exercise 1: Think about sentences that can be formed with these two verbs (examples below). Reflect on when and why you use *get* and *become*. What factors, if any, influence your decision?

a) *I want to get rich.*

b) *I want to become a surgeon.*

Exercise 2: Look at the following graphic. What observations on the use of *get/become + adjectives* can you make when looking at a corpus of spoken English?

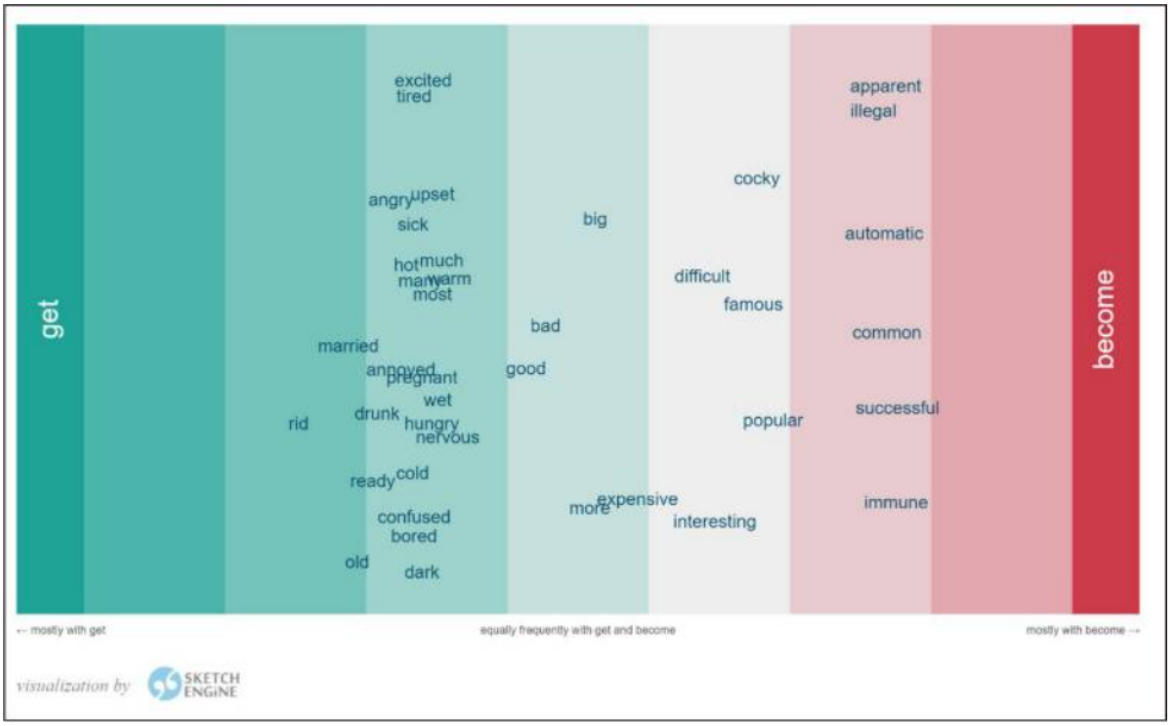

Figure 1. Get/Become + adjective in spoken English

Exercise 3: Now look at a graphic, showing collocations of *get/become* + nouns in a corpus of Web-English (written). What observations can you make here?

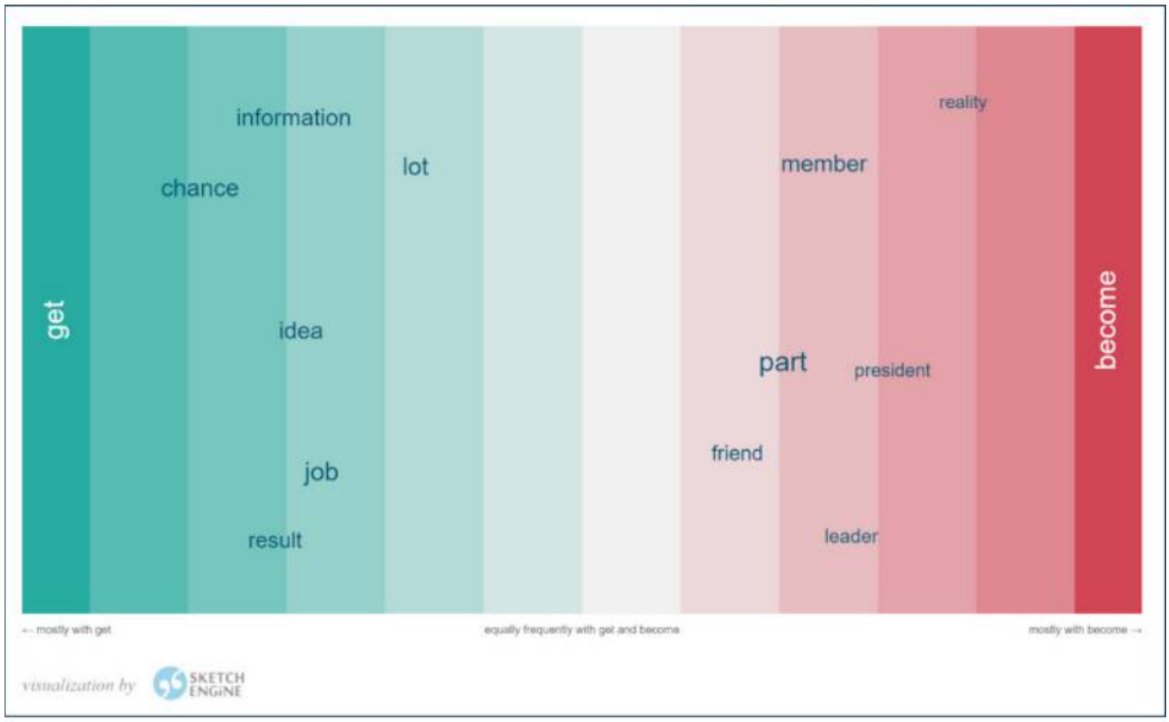

Figure 2. Get/become + noun in written English

Exercise 4: Before, you looked at *get/become* + adjective in spoken English and *get/become* + noun in written English. Now consult Figures 3 and 4 and investigate potential changes between written and spoken English.

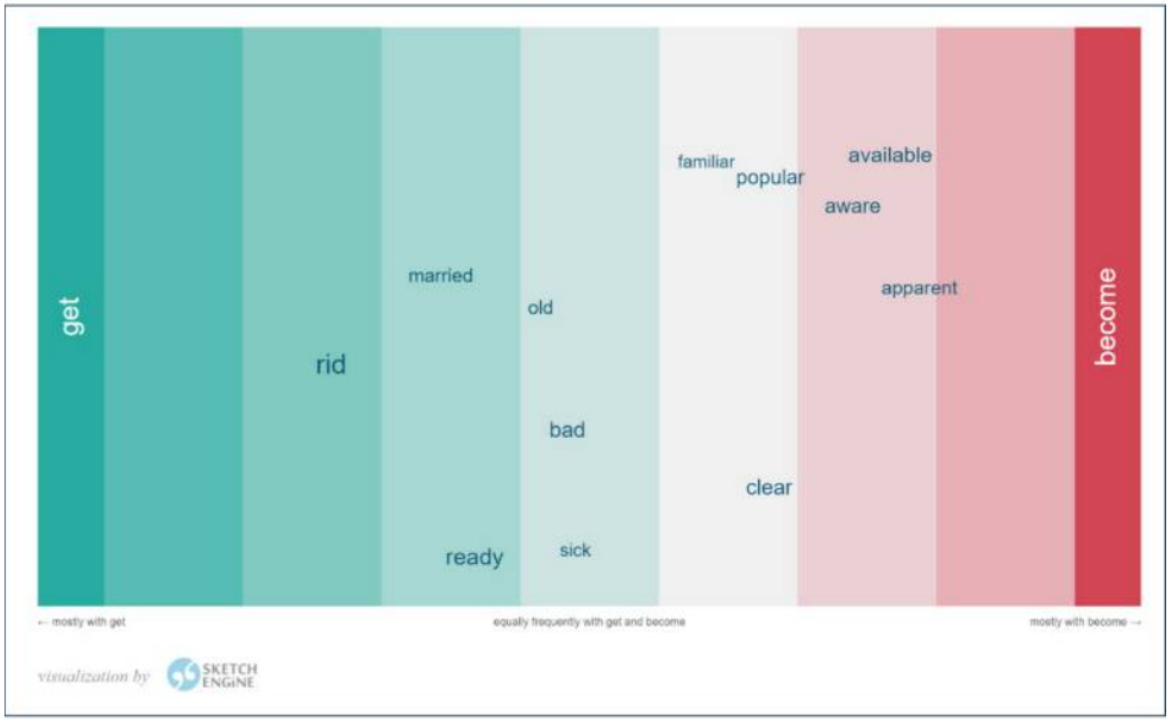

Figure 3. Get/become + adjective in written English

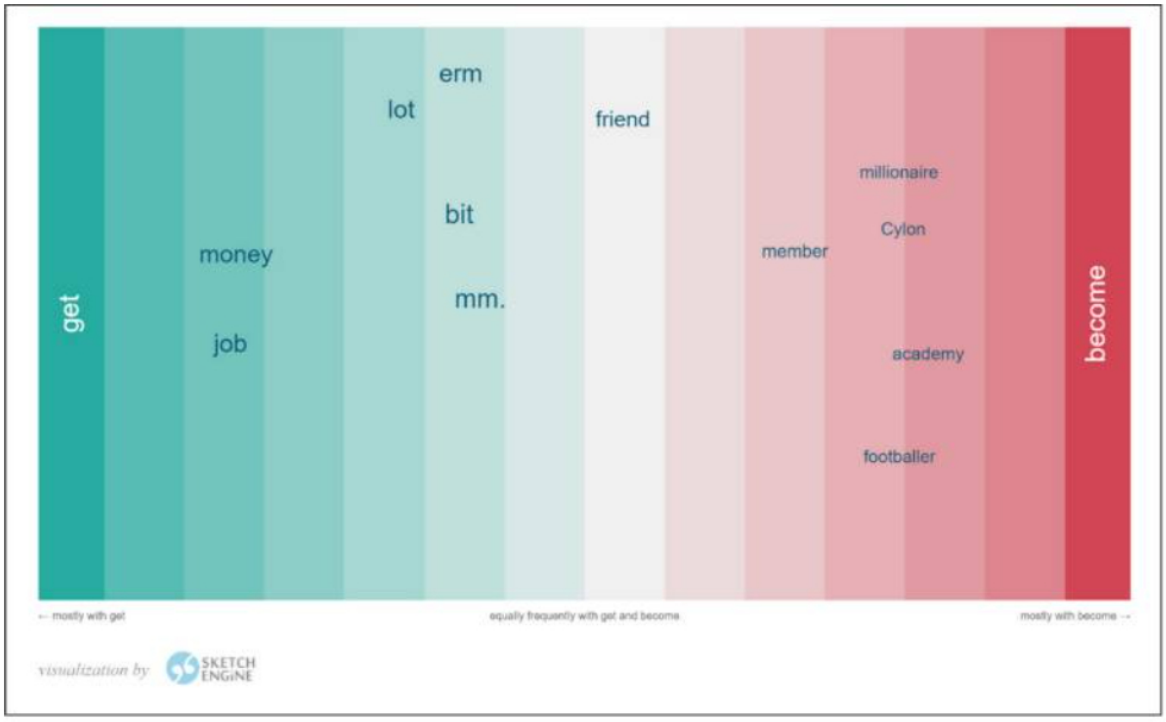

Figure 4. Get/become + noun in spoken English

Activity 3: Writing analysis using software

Vocabulary selection in class is sometimes based on unsystematic observations obtained from the teachers' experiences and the students' output. This can include bias and beliefs about what good language is, which can limit the efficacy of instructional practices (Arellano, 2022). Additionally, some teachers focus only on the problems rather than on the learners' progress as well. To tackle this issue, we suggest the following procedure to systematise the analysis from students' writing samples using corpus-based tools regarding teachers' feedback.

1) Download software to run a lexical analysis

There are various software tools you can use to analyse texts. In this section, we focus on AntConc (http://laurenceanthony.net/software/antconc/). This is a free quantitative software tool that analyses the most frequent words in a collection of texts and how these words are combined with others.

2) Prepare the data

All the students' writing samples must be converted into .TXT format—without images—so that they can be read easily by the software. After uploading these texts, go to the Keyword list tag and simply press Start. You will immediately find a list of words ranked by their frequency, as seen in Figure 5.

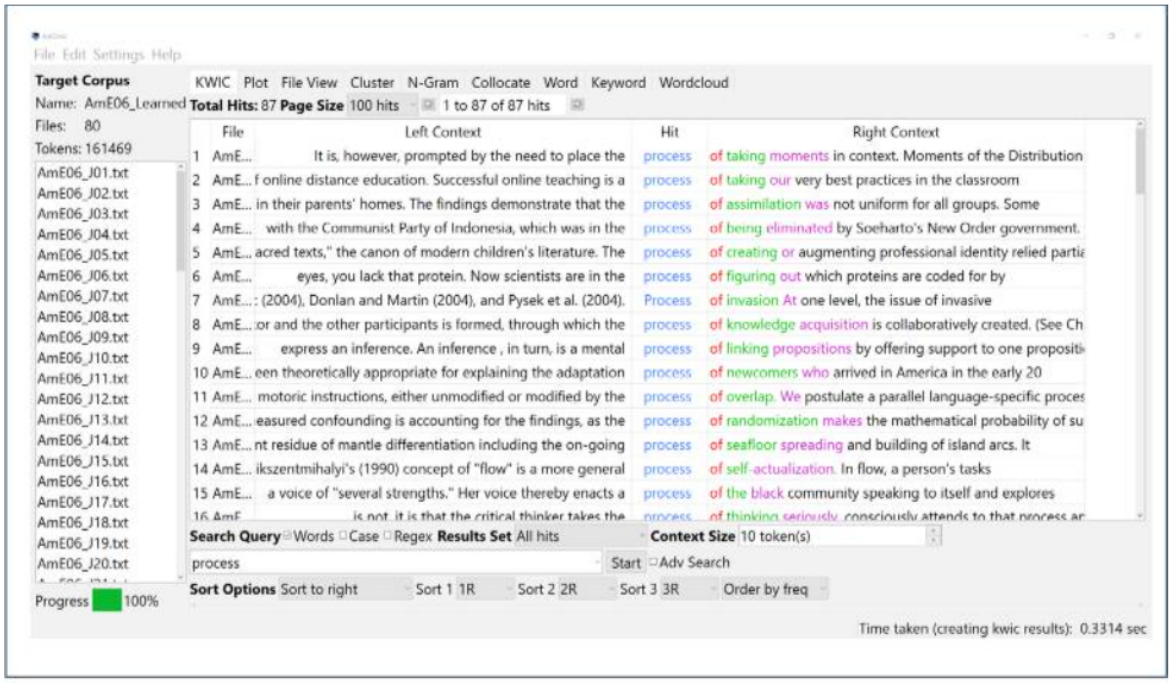

Figure 5. Keyword analysis in AntConc

3) Play with the data

You can explore the data in more detail by clicking any word to see examples of its usage in context. For instance, in Figure 6, the word 'work' is normally accompanied by prepositions, so you can use these concordances to illustrate their use in AntConc.

Figure 6. Examples of the word 'work' in AntConc

From here, we can explore the contexts and collocations of the keyword further. Whilst in *KWIC,* the list is already ordered according to the most frequent collocation to each side of the keyword, the Collocate tab provides us with information on the strongest collocations as well as some statistical information. For more contextualised input, the tab File view directs us to the actual document where the concordance line originates from, allowing the examination of keywords and collocations in their natural habitat.

4) Interpret the results

With access to these corpus-based tools, we can allow our students to autonomously investigate and explore language patterns. To start this analysis, students can consider the following range of questions:

- Which are the most frequent words?
- How are these words combined?
- Are the words taught in class used?
- Are there any words spelled incorrectly?

These questions are only one point of entry into student-led corpus analysis, and we invite every practitioner to expand this list as they see fit for their context.

Conclusion

Corpus linguistics has revolutionised language analysis. In particular, frequency analysis has been used to do research in social sciences and humanities, but it can also be employed as a technique to raise awareness about language use and to systematise feedback practices (O'Keeffe et al., 2007). This article has explored some functions of software packages, but teachers can enrich the analysis with other useful functionalities. Furthermore, the traditional numerical examination can be complemented with qualitative approaches, especially regarding collocations in context. Moreover, the options for visualisation offered by most software packages may make the results more visually appealing for teachers and learners. In this way, corpus linguistics and the use of analytical software enables students to conveniently and readily study vocabulary in context.

References

Arellano, R. (2022). *A discursive study of ideologies in an EFL teacher training program in Chile: The case of applied linguistics instruction.* [Doctoral dissertation, The University of New South Wales]. UNSW. https://doi.org/10.26190/unsworks/24397

Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes, 35*(1), 30-40. https://doi.org/10.1016/j.esp.2013.11.004

Gerigk, K. (2022). How to engage and motivate Generation Z in German EFL classes: A mixed-methods enquiry. *IATEFL RESIG Newsletter. 37*, 12-18. https://eprints.lancs.ac.uk/id/eprint/194563

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22*(3), 319-344. https://doi.org/10.1075/ijcl.22.3.02lov

McEnery, T., & Hardie, A. (2012). *Corpus linguistics.* Cambridge University Press.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching.* Cambridge University Press.

Timmis, I. (2015). *Corpus linguistics for ELT.* Routledge.

Rodrigo Arellano is a lecturer at Universidad de la Frontera (Chile) and The University of New South Wales where he completed his PhD in Applied Linguistics. He is interested in Discourse Analysis, Teacher Training, and Corpus Linguistics.

rodrigo.arellano@ufrontera.cl

Kevin Gerigk is a doctoral researcher and associate lecturer at Lancaster University. He works as a research associate at Aston University, and has taught in Germany, England, and Chile. He is interested in Data-Driven Learning, Corpus Linguistics, and TESOL.

k.gerigk@lancaster.ac.uk