Árboles y Rizomas Vol.V, Nº 2 (Julio-Diciembre, 2023): 144-150
Universidad de Santiago de Chile, ISSN 0719-9805
https://doi.org/10.35588/ayr.v5i2.6408

Review. CQPweb, BNClab, and CorpusMate and their applicability to the DDL classroom.

Introduction

Corpus Linguistics (hereafter CL) and data-driven learning (hereafter DDL) have started to find increased attention in the EFL (English as a Foreign Language) context over the past decade or so. With this increased popularity, the number of corpus analysis tools has multiplied. In order to offer some guidance and clarity on what tools may be most beneficial for data-driven EFL contexts, I discuss three commonly used corpus analysis tools in this review, in which I set out to briefly define the concept of CL and DDL for EFL contexts. Thereafter, I review three corpus analysis tools in terms of their suitability for DDL. I first review *CQPweb* which can be used by the teacher as a reference for materials design. It provides access to numerous corpora of written, spoken, and learner language. Then, I review *BNClab*, which allows a) to contrast contemporary spoken and written British English, and b) to contrast spoken British English from the early 1990s with that from the 2010s. Hence, learners can analyse language change and the differences between the written and spoken modes of English. Lastly, I review *Corpus Mate*, which combines various genres and topics into one large database. The corpus can be investigated as one whole dataset or be subdivided into topics. The user can also discriminate between spoken (e.g., TEDtalks) and written (e.g., Wikipedia) sources. This again is valuable to expose the learners to different genres, modes and target variations of English. All three analysis tools have their own affordances in the classroom and are suited to cover a distinct aspect that EFL teachers may find interesting to further explore in their teaching context.

Corpus Linguistics

CL refers to large collections of machine-readable texts (written and/or spoken) that have been collected and compiled in a principled manner to allow for linguistic analysis of patterns of lexis and grammar (McEnery & Hardie, 2012). These corpora (plural of corpus) can range from as little as 1m words, e.g., the BE21 (Baker, 2023), to as many as 52bn words, e.g. the EnTenTen21 (available on SketchEngine). The main advantage of using large collections of text in the classroom is that learners no longer solely rely on the intuition of materials designers, themselves, or the teacher. Instead, they can draw on a collection of language examples that have been produced in the real world with a communicative purpose. Whilst it lies still with the teacher to select and prepare the corpus-derived exercises (McCarthy & McCarten, 2022), which may add work and a new skillset to the teachers' responsibilities, it has been argued that the inclusion of corpus methods in teacher training modules has had a positive outcome on their professional development (Farr & O'Keeffe, 2019, p. 279). Hence, this review may help to increase accessibility to the concepts and options for corpus-assisted DDL applied to the classroom.

Data-driven learning

The use and analysis of corpora in language teaching is commonly referred to as DDL. As Crosthwaite et al. (2021, p. 1) state: "The use of corpora for the purposes of language teaching and learning [is] commonly known as 'data-driven learning' (DDL)". Although DDL requires some level of corpus literacy of the users (Templeton & Timmis, 2023), the benefits of using corpus-assisted DDL can be deemed considerable.

Using DDL in the EFL classroom has been described as a learner-centred, democratic approach that allows for noticing and peer scaffolding (Gilmore, 2015). By giving learners the opportunity to explore corpora and answer questions about language, the learners become Sherlock Holmes and conduct linguistic research (Johns, 1997, p. 101). Furthermore, using corpus data in a DDL lesson creates a unique opportunity to experience live language use in an EFL context (Templeton & Timmis, 2023), which learners otherwise might not have without travelling to the target country, since it may not be accessible to every learner.

Commonly, DDL can come in two distinct forms: direct and indirect. The latter referring to using CL to inform materials design is predominantly used by publishers, coursebook writers, and teachers for task design. Direct DDL refers to the use of corpus data and tools by teachers and learners directly in the classroom (Farr & Hagen Karlsen, 2022, p. 330). Whilst this review is not discussing specific exercises for classroom use, the interested reader may want to refer to Arellano and Gerigk (2023), who offer a set of tasks for language teaching and assessment, using a corpus-informed approach to DDL.

CQPweb

CQPweb (https://*CQPweb*.lancs.ac.uk/) is a freeware web-based corpus analysis tool (Hardie, 2012). The tool is slightly more advanced as it allows the user to conduct thorough language analysis. Hence, I argue that *CQPweb* may be used best as a resource to create teaching or reference materials used by the teacher. Whilst users need an account to access the corpora and analysis tools, a standard subscription is free and comes with access to a large host of corpora, such as the British National Corpus (XML edition) and learner corpora from different L1 backgrounds. The functions offered by *CQPweb* are those of standard corpus analysis tools: the users can run Keyword-in-Context[1] (KWIC) searches (Standard Query or Restricted Query), word frequencies (Frequency List[2]) and keyword analyses (Keywords[3]). In the Restricted Query, the user can set certain parameters in dependency on the metadata that is available for each corpus. This means that the user can, for example, restrict the results to young speakers (e.g., in the spoken BNC2014), which holds affordances for materials design for communication practices.

Firstly, it must be acknowledged that the interface, whilst visually simplistic, is extremely user-friendly and mostly self-explanatory. This should make it easy to access the functions of *CQPweb* even for novice users. *CQPweb* allows the teacher to download KWIC lines, which can be used in gap-fill exercises or to demonstrate the use and meaning of a word in context. Frequency Lists, on the other hand, allows the teacher to isolate lexical and grammatical words that the learners are likely to encounter when using English outside the classroom. Furthermore, word frequency can also lead to specialist vocabulary and make explicit technical terms, which are likely to reside in the lower frequency bands in a Frequency List. Moreover, Keywords can be used to identify topics that are representative of the time span when the corpus was built, and may illustrate associated vocabulary. For example, a Keyword search between the BE06 and the BE21 corpora in *CQPweb* makes clear that in 2021 the big topic was COVID-19 (see Figure 1 below). The first five keywords refer to the outbreak of the pandemic in the year 2020, the global lockdowns, and the efforts to find and roll out a

[1] KWIC shows the node word, i.e., the word we enter into the query, in centralised view. To the right and the left of the node word, we can see the context in which the node word occurs.
[2] Frequency Lists order the occurrence of all the words, lemmas or tokens in the corpus in rank of their frequency of occurrence from most to least frequent item.
[3] A keyword is a word that occurs statistically significantly more or less often in corpus A when compared to corpus B. This can help us identify genre or topic specific vocabulary and grammar.

vaccination programme. These words did not occur in the BE06 to that extent. In addition to already existing corpora in *CQPweb*, the user can also choose to upload and create their own corpora into the interface. This is a valuable feature for teachers who choose to analyse and contrast their learners' language production.

Figure 1. Keyword results (top 5) between BE06 and BE21. CQPweb (accessed 18 October 2023).

Keyword list for whole "British English 2006" compared to corpus "British English 2021";
using Log Ratio (with 0.01% significance filter, adjusted LL threshold = 36.2);
items must have minimum frequency 3 in list #1 and 3 in list #2.

|< << >> Choose action...

No.	Word	In whole "British English 2006":		In corpus "British English 2021":		+/-	Log Ratio	Log likelihood
		Frequency (absolute)	Frequency (per mill)	Frequency (absolute)	Frequency (per mill)			
1	2020	6	5.23	322	279.83	-	-5.74	393.86
2	vaccine	3	2.62	120	104.29	-	-5.32	141.95
3	lockdown	6	5.23	207	179.89	-	-5.1	240.01
4	Pandemic	9	7.85	259	225.08	-	-4.84	291.99
5	vaccination	3	2.62	64	55.62	-	-4.41	68.19

Whilst *CQPweb* comes with a lot of advantages, there are certain downsides to it as well. More advanced functions such as collocations, word distribution, and dispersion require some expert understanding and may not be accessible to novice users. These functions are hidden in a drop-down menu to the top left of the page of the KWIC view. Another consideration that users need to be aware of is that the default view of KWIC lines is ordered by file or text name. If users wish to examine a random sample of the language, this must be manually adjusted. However, these few limitations can be overcome very quickly with experience. The largest downside, however, is the different categories of licences, which may restrict access to certain corpora.

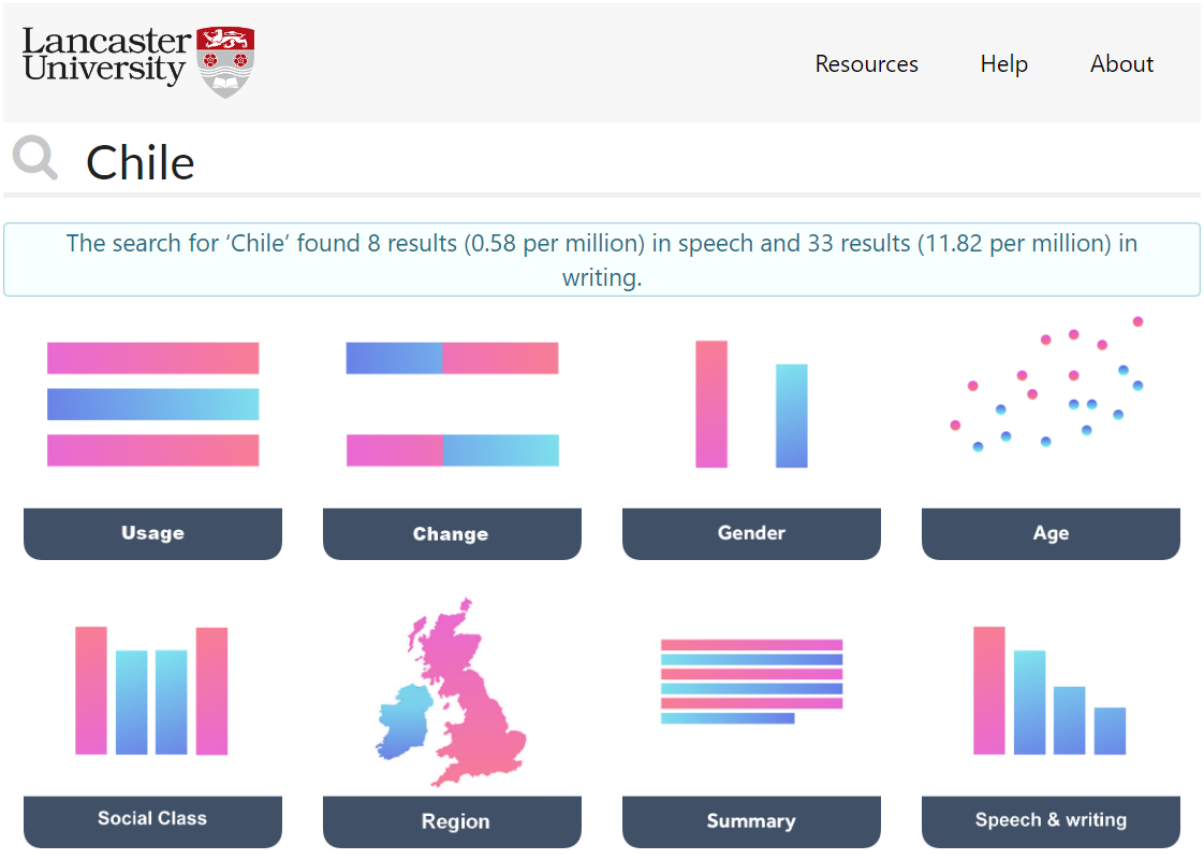
On the whole, it can be said that *CQPweb* is a fast, reliable, and user-friendly corpus analysis interface. It is suitable for novice users who are interested in the basic functions and query types around CL. More advanced features, however, may require training or more time to explore and make sense of them. The restriction of access to certain corpora may, indeed, pose a limitation to teaching and research; however, the sheer number of available corpora and genres on *CQPweb* almost certainly allows to find alternatives for the restricted corpora, and offers the opportunity to prepare materials from different genres that learners might find interesting and useful.

BNClab

BNClab (http://corpora.lancs.ac.uk/*BNClab*/search?query=,Brezina et al., 2018) is another web-based interface that can be used in a browser. *BNClab* is hosted and managed by Lancaster University. Whilst this tool does not enable the selection of different corpora, it uses the spoken and written part of the British National Corpus 2014 as a reference. *BNClab* is a user-friendly tool, which offers the opportunity to investigate and illustrate words and phrases from various aspects. The user can choose to look at KWIC lines to investigate the node word in context. Furthermore, *BNClab* allows the user to explore differences in usage of a word in dependency on Gender, Age, Social Class, and Region of the speakers, as shown in Figure 2. These

functions allow the learner to explore how people from different geographical or socio-economic backgrounds use English. In the classroom, this can be helpful to discover regional dialects or the type of English that is used by speakers of the same age as the learners, which may have an affordance for the development of an L2 sociolect. Perhaps the largest affordance of *BNClab* for the field of ELT is that it comes with a host of exercises, which are mostly freely available and are designed as DDL tasks in conjunction with *BNClab* (access: https://wp.lancs.ac.uk/corpusforschools/esl-teaching-materials/). The tasks refer to Current English Use, Spoken Communication, and Communicative Skills. Each task consists of corpus examples and is to be used in concert with the analysis interface of *BNClab*. The available teachers' guide is a handy supplement which gives invaluable tips and expert advice.

Figure 2. Homepage of *BNClab* with analysis options.



The main advantage of *BNClab* is that it is fully free and easily accessible. The interface itself is self-explanatory and easy to use, whilst it also gives the opportunity for more advanced research. The interface is relatively fast, which makes it a suitable tool that can be used in the classroom with little effort. *BNClab* comes with prepared exercises and a teachers' guide, which allows for an accessible DDL class with little preparation time for the teacher, be that an expert in the field of CL or a novice user of the approach.

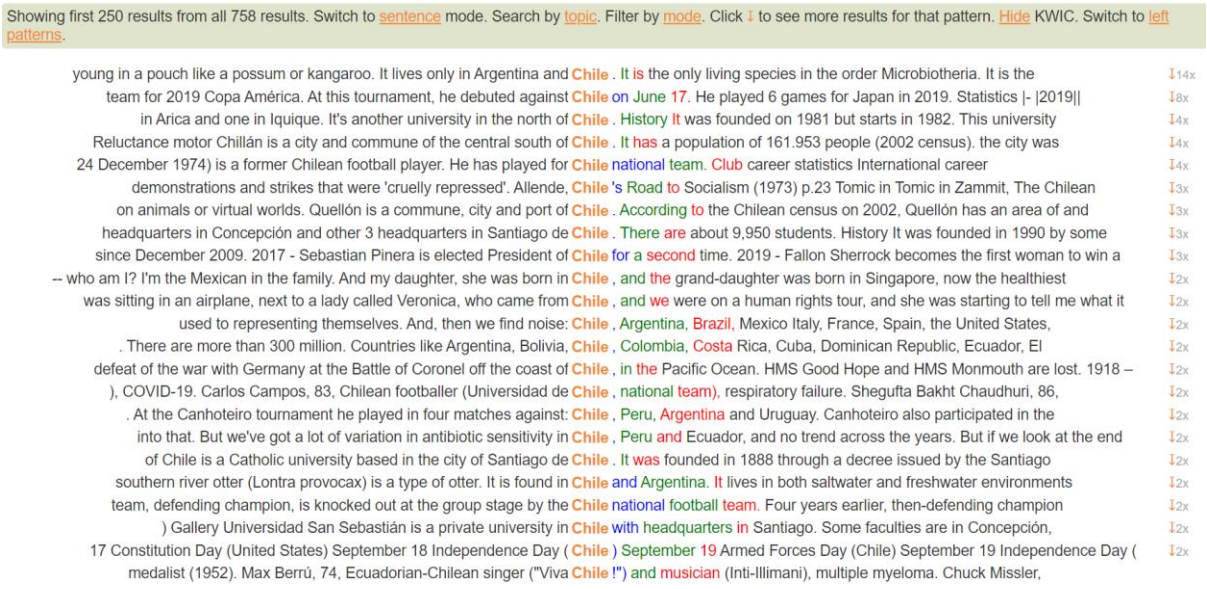
Whilst the advantages abound, it must be noted that *BNClab* is restricted to British English exclusively. This is because it is based on and accesses the British National Corpora of 1994 and 2014 for language analysis. If one is not particularly interested in the British variation, this tool may be less suitable.

Although limited to British English only, *BNClab* is an excellent resource for language teachers that wish to implement DDL into their classrooms at little cost or effort. The tool is user-friendly and time-saving as it comes with a whole host of EFL tasks that are readily and freely available through the companion website. In light of this, *BNClab* can be deemed a prime tool for DDL classrooms and gives more advanced users the opportunity to device their own exercises.

CorpusMate

CorpusMate (Crosthwaite & Baisa, 2023) is a web-based online interface, which gives access to a large collection of corpora of spoken and written English (n=around 50m words). The interface allows users to pre-select language from either spoken (e.g., TEDtalks) or written sources (e.g., Simple English Wikipedia), as well as from distinct topic areas such as History or Physics. This allows the learners to be exposed to different varieties of English as well as a diversity of genres from many different countries and sources around the globe. *CorpusMate* is particularly useful for teaching English to young learners as the dataset has been treated with a profanity filter to ensure that no bad language occurs in the results. Furthermore, the user has options to run various queries such as a wildcard to expand their searches[4]. All commands are explained and displayed in accessible language at the bottom of the homepage. Once the query is run, the user can choose to explore collocates of the node word, the distribution across various genres, n-grams, or more complex linguistic patterns. The latter is particularly beneficial in the exploration and teaching of collocations and phrasal verbs. This is because the corpus query returns language patterns, in which these phrases occur, as well as information on the frequency, i.e., the commonness, of their occurrence in the various genres. The default view, however, is a KWIC view, in which the node word is displayed together with its most frequent word to the right as shown in Figure 3 below.

Figure 3. KWIC view in *CorpusMate* with the strongest collocation to the right of the node word.



[4] A wildcard means that an asterisk replaces an individual letter or a combination of letters within a word, or a prefix or suffix. For example, the wildcard search *h*t* would return *hat, hot, heat,* and *hit.*

CorpusMate is a user-friendly, self-explanatory corpus analysis tool, which can be used in the DDL classroom with only little preparation and by novice teachers with only little corpus literacy. The results are displayed in a coherent and easily recognisable way, which eliminates distraction by avoiding complex queries and commands. Through this, the learners as well as the teachers are given the opportunity to explore raw language data, to compare words and phrases, to discover collocations and to contrast various genres and modes of communication, without having to spend time on learning complex query commands.

However, it must be acknowledged that *CorpusMate* may not lend itself to more complex analysis. This is because the user cannot, to my knowledge, access the source corpora individually. Furthermore, it appears that *CorpusMate* aims at displaying language patterns and occurrences across genres and varieties whilst omitting statistical information. This is not necessarily a limitation as the tool is designed for classroom use rather than advanced linguistic analysis.

Hence, it can be said that *CorpusMate* is an easy-to-use corpus analysis interface, which offers free access to various genres and topics. Whilst it can be seen as a limitation, I believe that the tools' simplicity makes it the most suitable analysis interface for the classroom. Learners and teachers are unlikely to be distracted or overwhelmed by an abundance of functions and statistics. Instead, they can focus on what should be the aim of a DDL lesson: language in use and in context.

Final appraisal

All three corpus analysis tools are a great addition to the classroom. They enable the implementation of DDL at a low cost and effort. The tools are freely available and allow for teacher creativity as well as for learner autonomy. This freedom in use allows learners to explore the English language in a semi-immersive context, which may increase the appropriateness of the language forms they produce in different contexts. Furthermore, the teachers are given more creativity in the design of materials. Teachers may also find it relieving that they do no longer need to rely on their intuition, which seems particularly daunting if English is an L2 for the teacher, too. Hence, corpus-assisted DDL holds many an affordance for EFL countries, where access to native speakers of English may be limited.

References

Arellano, R., & Gerigk, K. (2023). Insights from corpus linguistics: Using keywords-in-context to teach and assess English learning. *English Australia Journal*, *39*(2),66-75. https://search.informit.org/doi/10.3316/informit.383511093653714

Baker, P. (2023) A year to remember? Introducing the BE21 corpus and exploring recent part of speech tag change in British English. *International Journal of Corpus Linguistics*, *28*(3),407-429. https://doi.org/10.1075/ijcl.22007.bak

Brezina, V., Gablasova, D. & Reichelt, S. (2018). *BNClab*. http://corpora.lancs.ac.uk/*BNClab* [electronic resource]. Lancaster University.

Crosthwaite, P. & Baisa, V. (in press). A user-friendly corpus tool for disciplinary data-driven learning: Introducing CorpusMate. *International Journal of Corpus Linguistics*.

Crosthwaite, P., Luciana, L., & Wijaya, D. (2021). Exploring language teachers' lesson planning for corpus-based language teaching: A focus on developing TPACK for corpora and DDL. *Computer Assisted Language Learning*, *36*(7), 1-29.

Farr, F., & Hagen Karlsen, P. (2022). DDL Pedagogy, Participants, and Perspectives. In R.R. Jablonkai & E. Csomay (Eds.), *The Routledge Handbook of Corpora and English Language Teaching and Learning* (pp. 329-343). Routledge.

Farr, F., & O'Keeffe, A. (2019). Using corpus approaches in English language teacher education. In S. Walsh & S. Mann (Eds.), *The Routledge handbook of English language teacher education* (pp. 268-282). Routledge.

Gilmore, A. (2015). Research into practice: The influence of discourse studies on language descriptions and task design in published ELT materials. *Language Teaching*, *48*(4), 506-530.

Hardie, A. (2012). *CQPweb* – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, *17*(3), 380–409.

Johns, T. (1997) Kibbitzing one-to-ones (web notes). BALEAP: Academic Writing. University of Reading. http://www.eisu.bham.ac.uk/johnstf/pimnotes.htm

McCarthy, M., & McCarten, J. (2022). Writing corpus-informed materials. In J. Norton & H. Buchanan (eds.), *The Routledge handbook of materials development for language* teaching (pp.170-184). Routledge.

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

Templeton, J., & Timmis, I. (2023). A Flexible Framework for Integrating Data-Driven Learning. In K. Harrington & P. Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching* (pp. 39-58). Springer International Publishing.

Reviewed by

Kevin Frank Gerigk:
Mail: k.gerigk@lancaster.ac.uk ORCID: https://orcid.org/0000-0001-6187-2943
Lancaster University, United Kingdom.