Using corpora in language teaching, learning and use.
Alex Boulton, Corinne Landure

To cite this version:
Alex Boulton, Corinne Landure. Using corpora in language teaching, learning and use. Recherche et pratiques pédagogiques en langues de spécialité - Cahiers de l’APLIUT, Association des professeurs de langues des Instituts universitaires de technologie (APLIUT), 2016, 35 (2). hal-01850689

HAL Id: hal-01850689
https://hal.archives-ouvertes.fr/hal-01850689
Submitted on 6 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Using corpora in language teaching, learning and use

Alex Boulton

Alex Boulton est professeur en anglais et didactique des langues à l’Université de Lorraine. Il enseigne dans différents départements avec des cours de langue anglaise, de linguistique générale, de didactique, de linguistique de corpus et de méthodologie de la recherche. Ses intérêts de recherche concernent au plan large le lexique, l’enseignement/apprentissage des langues, l’exploitation des TIC ; et plus précisément l’apport des corpus (data-driven learning) en didactique des langues. Il a de nombreuses publications dans ces domaines et fait partie de comités de différentes associations dont l’AFLA (vice-président), le GERA S, TaLC et EuroCALL ; et de revues dont ReCALL (co-éditeur), Alsic, ASp, Eurocall Review, IJC AL LT, JALT-CALL Journal et Al-Lisi nyat. Il est responsable du Crapel (Didactique des langues et sociolinguistique), équipe de recherche à l’Atilf (UMR 7118 : Université de Lorraine et CNRS), et directeur adjoint de la première UFR Lansad en France. alex.boulton@univ-lorraine.fr

Corinne Landure

Corinne Landure enseigne au département de Techniques de Commercialisation à l’IUT d’Épinal, composante de l’Université de Lorraine. Docteure en linguistique anglaise, spécialité didactique des langues, ses recherches portent sur l’exploitation directe de corpus générique et spécifique par des apprenants Lansad. corinne.landure@univ-lorraine.fr

Language corpora have been shown to be of substantial help in improving learner writing, both for advanced students majoring in the language as well as for students with lower levels of proficiency needing language for specific purposes. This paper outlines two very different ways an online corpus has been successfully used as a reference resource by our learners. The first is in a distance learning context: the teacher highlights issues in the student’s academic writing and
suggestions corpus uses rather than simply giving the correction, the objective being to raise awareness of and sensitivity to the language problems identified. The second derives from a classroom situation where the learner himself chooses the items to check in a professional simulation writing task; the teacher guides the process, again with the intention of fostering the learner’s autonomy.

corpus, data-driven learning, reference resource, writing, guidance, language awareness, autonomy

L’exploitation de corpus pour l’enseignement, l’apprentissage et l’utilisation des langues


corpus, ressource de référence, guide, éveil linguistique, accompagnateur, autonomie

30 juin 2016

Introduction

Corpora can be used to study language in all its forms and uses. In language teaching and learning, one of its most common functions has been to inform dictionaries, grammar books, usage manuals, textbooks, syllabuses, tests, and other resources. Most of these keep corpora in the hands of expert corpus linguists, but there is little reason why end-users cannot also make use of such tools and techniques. Such approaches go by
different names, including corpus consultation and data-driven learning. While in reality both are used to refer to the same thing, there is potentially an enormous difference between consultation and learning; while one might assume that the act of looking up a word in a dictionary (consultation) might lead to long-term retention, there is no guarantee that this will be the case, even if it is the objective. Both the present authors have used corpora extensively for both purposes (e.g. Boulton 2016; Boulton & Tyne 2014; Landure et Boulton 2010), and propose some short examples here of some of the activities genuinely used with their students. The present focus is primarily on using corpora as a reference resource, in line with some of Tim Johns’ original uses, often referred to as ‘kibbitzing’, where the process was written up and made available to outside learners. A sample corpus kibitzers are still available at [http://lexical.net/TimJohns](http://lexical.net/TimJohns). In our set of two pedagogical notes, the focus is on French learners of English using similar techniques accessing the same corpus, but the first with a master’s student in a distance English degree, the second with a lower-level IUT student in class.

The corpus used is the Corpus of Contemporary American English (COCA), a collection of 450 million words available on line ([http://corpus.byu.edu/coca](http://corpus.byu.edu/coca)). While personal use requires free registration, any search conducted generates a distinct URL which can be passed on to others so that they can see the query performed and the results. The same website has a number of other corpora for different varieties of English, as well as corpora in Spanish and Portuguese. Corpora can be found for many other languages on the web, as well as for specific varieties of languages (e.g. law, medicine, etc.). What follows is merely an indication of the types of information that can easily be retrieved from a corpus and usefully passed on to students.

**Case study 1**

A Master’s student writing a short research paper in English on ‘erotic fiction’ produced the following text as part of a draft abstract, submitted to the teacher for comment. The text is relatively accurate in most respects, with the exception of *storie*. One can only assume that the student wasn’t using a spellchecker, or hadn’t switched it to English, otherwise it would be underlined in red on her screen. This is a simple example of the potential of **COCA underwent a facelift in May 2016 after the activities were conducted and the paper written up. The functions are the same and we have left the original screenshots for authenticity, but the new interface is cleaner, more intuitive, and works better on mobile devices.**
basic technology to help students in their L2 writing – but is in fact underexploited.

The rest of this short text presents few obvious orthographic, lexical or grammatical errors, but will immediately appear to be odd to most expert speakers of English. The teacher’s job here is not an easy one: s/he first has to identify a problem, then form an idea of what the learner wanted to say, and imagine how this could be better formulated. Then there is the question of what type of feedback to give, a subject which has generated enormous quantities of primary research as well as a number of meta-analyses (cf. Plonsky & Brown 2015). The main finding common to most is that feedback does help, but should not be too explicit. Systematically providing the (presumed) correct answers tends to lead to minimal cognitive processing and is quickly forgotten. Intriguingly, whether the teacher provides detailed indications (such as coding Prep for preposition, Tr for transitivity, etc.) or simply underlines the error does not seem to have much impact on the learners’ ability to correct their errors. There is a lesson here that we should not over-correct in terms of quality, but the same applies to quantity: it seems to be more effective to focus on a small number of errors that are considered important so that the learner doesn’t drown in too many errors of limited use for their present purposes. This is likely to be even more beneficial where the errors are related so that the learner has multiple opportunities to focus from different angles.

A final issue is the aim of the feedback itself. In the case of a CV, for example, the sole aim might be to help the learner improve her text as product. From a linguistic perspective, the emphasis might be on helping her to learn some new language for future use. In addition to these, we might even (or perhaps especially) want to help her improve the processes involved so that she can think about her future writing and consult resources successfully as the need arises.

In several places in this text, the teacher might not be able to find a dictionary or grammar reference that is explicit; such resources are wonderful generic tools, simplifying sometimes extremely complex points, but the goal being to deal with common issues they cannot, by definition, attempt to cater for every specific case. Where there is no obvious answer, the teacher might only be able to say ‘because’, a message that is difficult to transmit to the learner. This is where corpora can be helpful – not so much with errors as linguistic impossibilities, but with uses that are not ‘usual’.

What follows is not a generalisable “fiche pédagogique” in the usual sense that it is ready for use with other students, but a genuine example of the type of feedback that can be useful for L2 learners. In the present case, it
was fairly quick to produce, and made available electronically to the student concerned. An additional benefit, one might argue, is that it makes correction more interesting for the corrector.

This is the original text of the Master’ student mentioned above (reproduced with permission):

This successful genre suffers high criticism while remaining a world-wide best-seller phenomenon. This paper tries to put a finger on the reason of this conflict...

Most of all, the store's proximity to reality gives the impression to be the main reason of the heavy success amongst female readership.

The following feedback was provided on the following elements only using MS Word comments. Other aspects may be debatable.

| High criticism. This is where a corpus can be useful for collocation, as we don’t often say ‘high criticism’; try this and see if there’s anything that satisfies you: http://corpus.byu.edu/coca/?c=coca&q=43356224 |
| Put a finger on. Similarly, look at the distribution of this expression: http://corpus.byu.edu/coca/?c=coca&q=43356250 |
| Reason of. Check preposition: http://corpus.byu.edu/coca/?c=coca&q=43356267 |
| Gives the impression to be. For how to use this, look at what comes after this phrase here: http://corpus.byu.edu/coca/?c=coca&q=43356308 |
| Reason of. See above. |
| Heavy success. Another collocation issue; can you formulate a query similar to that for ‘high criticism” above? |

“High criticism”

The student clicks on the URL to see immediately what the query was (on the left in Figure 1), i.e. what adjectives occur frequently before criticism.
There is no need to learn the codes for different parts of speech; these can be selected from the drop-down POS menu. While various functions are available, this is a simple query using entirely default settings for presentation (list) from the entire corpus (as opposed to defined sections such as academic). In case of doubt, the user can click on the question mark on the right for additional help. Figure 1 gives the first 20 results in rank order with the number of occurrences in the corpus on the right. Most of these have meanings and uses which are not relevant to this particular case, but the student should notice a number of possible options for her purposes: harsh criticism, sharp criticism, intense criticism, widespread criticism, heavy criticism, strong criticism, all of which occur over 50 times in this corpus. If hesitating between them, clicking on the term to see examples in context, along with information about the source text. The student could also formulate her own query for high criticism and receive the message: “Sorry, there are no matching records”. In other words, for these two words which individually occur 209,144 and 16,848 times respectively, there is not one single attested use of them in this combination in a corpus equivalent to 7,000 books or 60,000 research articles. Such features highlight the importance of collocation (i.e. co-occurrence) in the ‘naturalness’ of language use.
“Put a finger on”

The URL provided this time shows the distribution of the phrase in the different sub-sections of COCA using the chart function at the top (Figure 2), a function which can be extremely useful when dealing with languages for specific purposes.

![Figure 2: Distribution of put a finger on](image)

Clearly there is nothing wrong with this phrase in itself, but it is very unlikely to occur in academic texts such as the student is attempting to produce here – on average only 0.03 occurrences per million words. The student can click on the column headings (e.g. fiction) to see distribution in sub-sections (general books, sci-fi/fantasy, etc.), the bars to see the occurrences in context. Awareness of genre and text type are crucial.

“Reason of”

Prepositions are a common source of difficulty, and here is perhaps the closest we get to a language ‘error’ as opposed to a usage issue. As figure 3 illustrates, the query uses an asterisk to stand for any word: *the reason*
This is quicker than choosing the POS tag for adjective, although there is correspondingly more ‘noise’. From the first 10 results above, the most frequent preposition by far is for with 2915 occurrences; the user can scroll down to find the reason of with only 68 occurrences (i.e. over 40 times less frequent), and click on it to see it in context – most involving different uses (e.g. the reason of course) or meanings (e.g. the reason of the law).

“Gives the impression to be”

The URL shows there are 48 occurrences of gives the impression; clicking on this provides short contexts, the first 10 of which are given in Figure 4 below. Of these, 14 are followed by a that-clause; the remaining 6 are followed by of plus a noun (lines 9, 12, 13, 17) or, in two cases, Ving (lines 7, 19).

Figure 4. Contexts for gives the impression
This presentation encourages the student to look at usage in context and detect recurring patterns – an extremely useful skill if it can be drawn upon in other situations. The particular advantage here is the presence of multiple examples rather than occasional serendipitous encounters with isolated occurrences which are likely to go unnoticed. Possible supplementary searches show that there are only 12 occurrences of \([\textit{give}] \text{ the impression to}\) (where the square brackets indicate any form of the word – give, gives, gave, given), and in 11 of those cases to is a preposition (e.g. \textit{I didn’t want to give the impression to} anybody that I was guilty of anything).)

“Heavy success”

Having seen various queries and the results they generate, the student might be in a position to use the corpus on her own. \textit{Heavy success} gives no hits at all; using an almost identical query to \#1, it should be easy enough to select adjective from the drop-down POS menu or input \([*]\) followed by success, as in Figure 5. As before, some of the results are clearly irrelevant, but there are various frequent possibilities to choose from, and which can be seen in context in case of doubt. Having the student look at or perform several queries of the same type is more likely to lead to uptake and continued, autonomous use in the future.

![Figure 5. Results for \textit{adj+criticism}](image-url)
Case study 2

As in the previous case study, we look here at an actual case corpus consultation that was seen to be useful for an EFL learner, especially with regard to usual, normal, natural use of language in context. In this case, however, we are dealing with a student in the second year of a DUT Techniques de Commercialisation (a two-year post A-level degree in Business and Marketing), i.e. not majoring in English and with relatively low level of proficiency in the language – approximately A2. Of course, this experiment can apply to all fields of studies. The objective here was to use the same online corpus (COCA) to help improve the quality of a short professional text. While the query types are largely similar, the problem areas were not indicated by the teacher, and it was up to the students to choose the points to look up in the corpus. The teacher’s role was largely reactive with respect to individual queries. This type of feedback was designed to help students improve their corpus consultation techniques, suggesting alternative search functions or ways of formulating their questions or more efficient queries including semantic searches, genre-sensitive queries, diachronic variations, etc.

The corpus was initially presented to the whole class, but students then used the corpus individually for their own questions. Such individualisation represents a substantial benefit of corpus consultation as each member of the class can work at their own pace on points that are relevant to them, a potential source of motivation. Students can also work in pairs or small groups, with the stronger and weaker students each benefiting from the others’ input not just on language, but on the interface and the techniques involved, analysing and sorting the results, etc.

Where students are allowed the considerable freedom they are here, there may be initial questions focus on the content and the objective – what are we doing and why are we doing it like this? Such questions are an important step on the road to more autonomous language learning. In practice, the first step is to choose a word or phrase to check, the actual question being secondary; in other words, the nature of the problem and types of possible improvements emerge from the search itself. This often means a cycle of several searches, some of which may additionally involve dictionary use or other resources, until such time as the student is satisfied they have managed to find relevant data and interpret it appropriately for their specific contextual use.
This case study examines the procedures used for the sentence below, produced by a student from this class, following the introduction to corpus use and a few tentative searches in previous lessons:

**These information are useful. I need these quickly on Monday 5 January.**

The student had a doubt about the use of *information* in this context. He has been warned about it before and knows it is ‘uncountable’, but such abstract knowledge is difficult to fully grasp and not always easy to translate into appropriate usage when the time comes. His first search is simply for information, but using the KWIC (keyword in context) presentation in COCA where the target item is centred and surrounding items colour-coded for part of speech (students should be aware though that the POS-tagging is automated and may contain some errors). Here is a sample:

![Figure 6. KWIC search for information](image)

Even in the random selection of 100 lines, there are no examples of *these* immediately to the left of *information*, nor *are* to the right. While corpus linguists are fond of pointing out that the absence of evidence is not evidence of absence, they are equally fond of stating that they are not interested in what’s theoretically possible but in what is actually probable. In other words, absence or rarity of a particular form in a large, well-constructed corpus may be taken to imply that it is at best relatively unusual, where the learner is typically interested in normal usage.

The student followed up this search by clicking the drop-down POS List menu and selecting *det.ALL* for any determiner, followed by a space and then *information* – the first 12 are given in Figure 7.
This lists the determiners occurring immediately to the left of information in descending order of frequency; since this information is in second place with 4339 occurrences while these information is not to be found anywhere near the top of the list, there is a clear reason to prefer the former, all things being equal.

A similar procedure allows a quick search for verbs following information (Figure 8). Here the student noted information are in 9th position with 304 hits (i.e. a frequency of 0.68 occurrences per million words), and decided to go no further. Fortunately, the teacher was on hand and managed to encourage the student to pursue the query.
Admittedly, the results suggest that *information are* is a possible string, but context is needed to see whether it a coherent syntagmatic unit (which it is not). Clicking on *information are* brings up concordance lines which clearly show that *information* on its own is not the subject of the verb (a random sample is shown in Figure 9).

**Conclusion**

The last example highlights that interpreting the results can be the trickiest phase in corpus consultation, requiring an ability to take a step back and think quite hard on occasion. On the one hand, one might think that this adds to the cognitive load; but on the other, it is crucial that it is relevant cognitive activity, and precisely the type of mental effort which is required for successful language learning. Similarly, such uses of corpus can be quite time-consuming – especially at first, though they speed up rapidly with regular use. But it is the process itself that is important, not the answer to a particular question. A learner who is capable of noticing reasoning language in this way, noticing usage in context and making sense of authentic data, is a more autonomous and sophisticated learner better equipped to deal with genuine future needs.


