How to use a corpus:

A teacher's guide to looking for patterns

Accessing a corpus can feel time-consuming and difficult. Is it really worth it? Can't I just use ready-made teaching activities with the answers included? Er, yes of course! But there may be times when 'the' answer is not satisfactory for you, or your students. A corpus can be one of the many tools in our teaching repertoire. But searching through millions of words won't just give us answers. We need to examine them and look for patterns that will lead us to the kinds of answers we need.

01

Remember you are always looking FOR something. Not at, with, or after. Have a purpose. You will probably find it more helpful if you focus on a topic related to your and/or your students' English language learning. Reasons for corpus exploration can come from your everyday experiences. For example:

- Your own language-related doubts and worries as you prepare your lessons, i.e. the items you check in a dictionary or a grammar, ask a colleague or look up in a search engine.
- Students' questions (we've all been asked questions to which we didn't know the answer!)
- Issues you identify while marking your students' work – both the ones that you're certain about and the ones you do not know how to fix.

02

Consider what you would like to learn more about. And yes, it will be you doing the learning. Corpus exploration typically starts with a specific search. For example:

a. a part of a word (e.g. which words end with '-able'?)
b. a word (e.g. how frequent is 'morrow'?)
c. sequences of words (e.g. how is 'pale sky' used?)
d. potential synonyms (e.g. what are the potential synonyms of 'help'?)
e. word classes (e.g. in which register are subordinating conjunctions most frequently found?)

The last two types of search may not be possible in all corpora because they require the corpora to have been annotated, meaning that they have been tagged semantically (d) and grammatically (e).

03

Explore. Keep in mind that most corpus investigations will be exploratory. You are trying to come to grips with how language is used. For this reason, give yourself some time to discover and be flexible with your search. Allow your reactions to what you see in the results to take you in different directions, if that is needed and/or relevant.

04

Reflect on the quantitative results first. This is a useful start to pattern finding. Corpus research is by no means limited to counting how many instances of a search item have been found, but frequencies are a good initial indicator of how language is used.

Example

Take the exploration of the difference between the verbs 'begin', 'commence' and 'start'. A first look at the frequencies tells you quite a lot. In the British National Corpus (BNC), there are 41,566 instances of 'begin'; 1,511 of 'commence'; and 39,316 of 'start'. These results (perhaps unsurprisingly) indicate that both 'start' and 'begin' are used at least 27 times more frequently than 'commence', and there seems to be a slight preference for 'start' over 'begin'.

@AislingOBoyle1 @vanderviana

How to use a corpus:

A teacher's guide to looking for patterns

05 **Go further and investigate** the frequency of the different forms of each of these three verbs in #4 (i.e. started, starts, starting). Looking at the different forms of the verb in the BNC, you can see that for all three verbs, 'begin', 'start', 'commence', it is actually the past forms – 'began', 'commenced' and 'started' – which are most frequently used. Our exploration is giving us evidence of how the words are normally used, or rather, their patterns of use.

06 **Examine some of the concordance lines** (i.e. short snippets of text around a specific search word/expression). By looking at instances of the word in context, we can begin to think about how it is being used, and for what purpose.

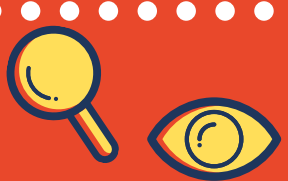
"I began to take notice of her when she was hardly thirteen years old."
"Criminal proceedings had been commenced against the managers of the Swiss company in Lugano"
"But it all started long before that."

The examples above from the BNC show that these verbs are being used to refer to a specific action that took place at a time prior to its reporting.

07 **Check the distribution** of your search word/expression across different genres or registers, if that breakdown is possible in the corpus website that you're working with. General corpora (i.e. those aimed at representing a language) will have texts in different categories (e.g. face-to-face conversation, newspaper texts, broadcast recordings, soap operas, novels), and they are likely to provide quantitative frequency breakdowns for your search item.
For example, the BNC interface provides breakdowns for seven sections: spoken, fiction, magazine, newspaper, non-academic, academic and miscellaneous.

A chart search for our exploration in #4 of 'begin', 'commence' and 'start' shows that these verbs vary in frequency across different corpus sections. 'Begin' is most frequently found in fiction; 'commence' appears more often in miscellaneous; and 'start' is used most often in spoken.
Even straightforward quantitative results already tell us a lot about the overall use of these three verbs (see # 4), the forms in which they are most commonly conjugated (see # 5), and where they appear most frequently (see #7). Hoorah! Patterns identified!

08 **Now verify.** Repeat the same search in another corpus to check whether the results are consistent. As a language teacher, you might be familiar with the boxes in learners' dictionaries, drawing attention to the differences between, for example, American and British English.
As #4 and #5 focused on British English, you could investigate the same three words in the Contemporary Corpus of American English (https://www.english-corpora.org/coca/) to see what you learn about them in another national variety.
Go ahead, give it a try.
Some answers at the end of the guide!



@AislingOBoyle1 @vanderviana

How to use a corpus:

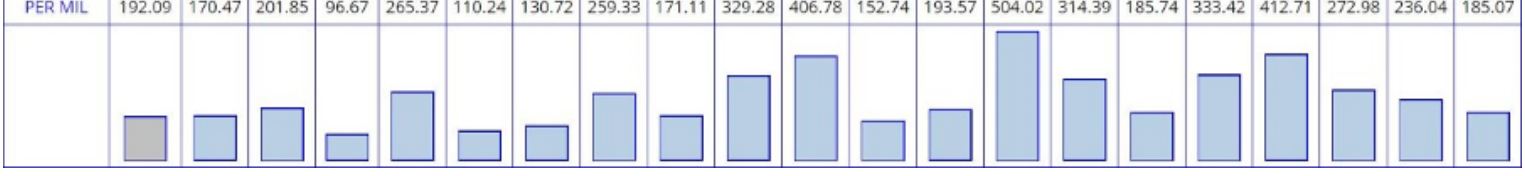
A teacher's guide to looking for patterns

09 **Consider the differences across English language varieties** beyond American and British. This will help you to raise your students' awareness of the fact that English is not monolithic - there are varieties of English and Englishes. The Corpus of Global Web-Based English (GloWbE / https://www.english-corpora.org/glowbe/) is an excellent resource in this regard, containing data from 20 countries. It is only made up of texts found on the web, so bear that in mind for teaching purposes. In #4-6, the focus of our exploration was lexical: the use of three different verbs with similar meanings. We can also investigate patterns of use in relation to grammatical items.

Example

So let's take carry out another corpus exploration, this time with a grammatical focus using GloWbE. The figure below shows the distribution of 'shall' as a modal verb. As in #4, reflect on the quantitative results first. They indicate differences across English varieties: e.g. Philippine English seems to employ this verb five times more than British English. Consider further exploration using #5-7.

SECTION	ALL	US	CA	GB	IE	AU	NZ	IN	LK	PK	BD	SG	MY	PH	HK	ZA	NG	GH	KE	TZ	JM
FREQ	364972	65941	27202	37472	26810	16338	10639	25007	7971	16914	16062	6564	8061	21798	12717	8426	14219	16000	11209	8299	7323
WORDS (M)	1900	386.8	134.8	387.6	101.0	148.2	81.4	96.4	46.6	51.4	39.5	43.0	41.6	43.2	40.5	45.4	42.6	38.8	41.1	35.2	39.6
PER MIL	192.09	170.47	201.85	96.67	265.37	110.24	130.72	259.33	171.11	329.28	406.78	152.74	193.57	504.02	314.39	185.74	333.42	412.71	272.98	236.04	185.07



US (United States), CA (Canada), GB (Great Britain), IE (Ireland), AU (Australia), NZ (New Zealand), IN (India), LK (Sri Lanka), PK (Pakistan), BD (Bangladesh), SG (Singapore), MY (Malaysia), PH (Philippines), HK (Hong Kong), ZA (South Africa), NG (Nigeria), GH (Ghana), KE (Kenia), TZ (Tanzania), JM (Jamaica).

10 **Investigate more patterns of use.** Consider if there any words which are recurrently used together with your search word. Corpus research shows us that words are not used randomly nor in isolation. Mostly, we make consistent language choices. We say or write sequences of words that we've been in contact with previously through our experience in the world, and we tend to stick with those choices. So, thinking about our exploration of 'shall', what choices do language users tend to make with the modal 'shall' in GB that are similar or different in the Philippines? **Remember #5: Go further and investigate.** Investigating the lexical verbs which collocate with 'shall' in the web-based texts from these two countries gives us the following results.

#	Great Britain	Philippines
1	See	Take
2	Say	Made
3	Take	Deemed
4	Go	Imposed
5	Make	Include
6	Give	Considered
7	Find	Apply
8	Remain	See
9	Come	Allowed
10	Know	Entitled

This shows us that 'shall see' is the top choice in GB, 'shall take' is the top choice in the Philippines and that there is limited convergence: just 'see' and 'take' appear in both lists in the same form. Look at the verb 'make', which is featured in both lists but in different voices. It is used in the active voice in GB while it appears in the passive voice in the Philippines. This actually reveals a main difference in the use of 'shall' in these two countries: all the top 10 lexical verbs in Great Britain are in the active voice whereas the passive voice appears in six results in the Philippines. Again, **remember #6: Examine some of the concordance lines** in order to develop a fuller understanding of how 'shall' is used in these two countries.

@AislingOBoyle1 @vanderviana

How to use a corpus:

A teacher's guide to looking for patterns

10 (c'td)

Great Britain / active voice	Philippines / passive voice
"As we *shall see*, it provides significant benefits which can offset the effects of ageing."	"Training for work with BECs *shall be made* part of seminary formation."
"I *shall say* some special prayers for them tonight."	"If there is any doubt, such risk *shall be deemed* to exist."
"In the name of all competitors, I promise that we *shall take* part in these Olympic Games"	However, after the 20th, a penalty *shall be imposed* for late payments."
"If you go to the city, no one *shall go* with you."	"The following criteria *shall be considered* in the conferment of awards:"
"The Arbitral Tribunal *shall make* its award in writing"	"Only one vehicle *shall be allowed* per family"

This investigation of 'shall' shows that corpus investigations will blur the usual pedagogical distinctions among vocabulary, grammar, pragmatics and so on. While our starting point was to check the distribution of a modal verb, the results revealed much more than we had initially in mind. This is in line with our use of language and why often 'the answer' is not enough. We, as language users, rely on our knowledge of language systems in an integrated way to make meaning. In other words, we do not think about, for instance, vocabulary and grammar separately before speaking or writing something.

And finally . . . ALWAYS compliment yourself on the work you have conducted, whatever your outcome.

It is not always easy to identify patterns in language use; indeed it might not always be apparent or possible. Practice is of utmost importance. The more you practise, the easier it will be for you to spot these patterns. Even if a pattern cannot be observed, your investigative work and time is never wasted for two very good reasons: (a) not having any differences is a result in itself, and (b) undertaking the kind of language investigation described here helps us to develop highly important analytical skills.

*****Spoiler Alert: Some answers to #8*****

The quantitative results show that there is a clear distinction between 'begin' (412,878 occurrences) and 'start' (578,246 occurrences) on the one hand, and 'commence' (4,672 occurrences) on the other. What we observe in the COCA is that 'start' appears more often than 'begin', which differs from the results in the BNC. This difference is probably related to the composition of these two corpora.

In relation to forms, the COCA results indicate that the three lexical verbs are most frequently used in the past: 'began', 'commenced' and 'started'. This reinforces what had been noted in the BNC.

With regard to sections, frequency counts confirm that 'begin' is used mostly in fiction and that 'start' appears most often in spoken. Because there is more spoken data in the COCA than in the BNC, this might explain the reason why 'start' is overall more frequent than 'begin' in the COCA when compared to the BNC results. 'Commence' in the COCA is most frequently found on the web. It is important to note that the sections in the BNC (spoken, fiction, magazine, newspaper, non-academic, academic and miscellaneous) do not match those in the COCA (blog, web, TV/movies, spoken, fiction, magazine, newspaper, academic).

The interface for the COCA also provides breakdowns for time periods: 1990-94, 1995-99, 2000-04, 2005-09, 2010-14 and 2015-19. We can therefore see how language use varies across different 5-year periods. The quantitative results seem to suggest that 'start' increased until 2010-14 and that 'begin' is decreasing from 2000-04. However, these provisional findings would need to be investigated further, especially in light of their diverse distributions across sections.

@AislingOBoyle1 @vanderviana