Digital corpora
language teaching and learning in the age of big data

Matt Absalom¹

<table>
<thead>
<tr>
<th>Potential impact</th>
<th>medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timescale</td>
<td>ongoing</td>
</tr>
<tr>
<td>Keywords</td>
<td>digital corpora, corpus linguistics, data-driven learning, language in context</td>
</tr>
</tbody>
</table>

What is it?

Using corpora to teach languages is nothing new and, while the term corpus linguistics hails from the 1940s, most language learning before the 20th century adopted a corpus approach – using a series of texts in the language under study as a type of corpus on which to base acquisition. With the advent of widespread computing in the latter half of the 20th century, corpora began to be digitised, rendering interrogation of large amounts of data a much simpler and more appealing prospect. Today, languages in all forms (written, spoken, performed, formal, informal, etc.) are captured all the time through online and digital platforms, apps, etc. meaning that the wealth of language data literally at our fingertips is enormous. This has triggered the development of appropriate tools to explore these vast data sets.

For language teaching and learning the possibilities fall into two categories: using existing corpora or creating your own corpora. A good place to start exploring language corpora is Sketch Engine (https://www.sketchengine.eu/corpora-and-languages/). You can sign up for a free 30 day trial and access all functions, featured corpora for all languages, as well as the corpus building capacities.

¹. University of Melbourne, Melbourne, Australia; mabsalom@unimelb.edu.au; https://orcid.org/0000-0003-3539-8832


© 2021 Matt Absalom (CC BY)
Which leads to the second type of activity: creating corpora. Apart from Sketch Engine, another relatively accessible option is #LancsBox (http://corpora.lancs.ac.uk/lancsbox/) which allows you to either interact with existing corpora or create your own.

Why use corpora? Applying corpora in your teaching and learning can support activities which involve inductive learning: analysing language to work out how something works, particularly in context. Utilising digital corpora, either those already available or creating your own customised corpora, streamlines this process as you can instantaneously produce all instances of, say, a particular grammatical feature or see how a word is used. You can also apply this to text types or genres – for instance, what do newspaper articles do that is different to short stories or how do people make doctor’s appointments over the phone compared to making a hair appointment? Many online language sites take a corpus approach such as Reverso Context (https://context.reverso.net/translation/).

**Example**

A constant stumbling block for learners of Italian is the choice of preposition. This often comes from the simplistic one-to-one translations presented in language textbooks, manuals, etc. In order to sensitise students to the importance of context in the correct selection of prepositions, I devised an exercise which used a small corpus created from the two literary texts that were under study at the time – I thought this would be useful pedagogically, since the students were already reading these texts and therefore would approach the task with less anxiety and more familiarity. I imported the texts into #LancsBox to create the corpus and then created lists of concordances (Figure 1) which showed the prepositions, *a*, *di*, and *da* in context. I gave students a table (Figure 2) to complete which helped guide their mining of the data. Essentially, they had to transpose the occurrences of the preposition from the original concordance lists of contexts from the texts in question into columns which showed the diverse functions of the prepositions: e.g. locative, genitive, introducing an infinitive, etc.
Figure 1. Example of list of concordances of preposition di

Verbi che vogliono di quando seguiti da un infinito
Per es.: ha finito di mangiare

Espressioni che vogliono digiando seguite da un infinito
Per es.: sono capace di farlo

Espressioni idiomatiche con di
Per es.: di buon'ora

Uso possessivo
Per es.: il libro di lei

Altri usi
Da definire

Figure 2. Example of table for students to complete: la preposizione di
Benefits

Existing language corpora provide endless examples of language in context in diverse registers, genres, time periods, and text dimensions. While predominantly text-based, there are also corpora of recorded language whether spontaneous, televised/broadcast, or scripted. Importantly, the work of constructing these language banks has already been done (and continues).

For those with developed Information Technology (IT) literacy, corpora tools offer a lot of scope for exploration of language and data-driven learning. Teachers can custom-build their own corpora or customise existing corpora. Students too can be instructed to use corpora tools to investigate how language works through accessing large arrays of exemplar texts.

Potential issues

The most glaring issue with digital corpora is technology. Corpus linguistics is the province of computer scientists and linguists and, while software tools are becoming more user friendly, building and interrogating corpora still require a significant effort even for those with reasonable IT skills.

In the example above, I decided to avoid wrestling with students’ capacity to use the software to access the corpus and provide them with an excerpt myself. This was largely because the year before this I had asked the previous group of students to download software, read the manual, load the corpus, and then carry out various tasks which remained beyond the majority of my students. The focus of my class was not corpus linguistics, this was simply a different way to approach the study of Italian so, in some respects, it is too much to expect language students to (want to) learn how to use digital corpora. Additional issues relate to accessibility of digital corpora which might be problematic for students with learning or physical disabilities, or limited access to technology. Finally, not all languages have the same number or variety of corpora readily available online.
Looking to the future

There is no doubt that we will continue to amass massive amounts of (language) data. It is also the case that digital corpora will lead to more nuanced development of translation and AI-supported language tools. Taking advantage of these developments and accessing digital corpora to support language learning, both in and outside formal settings, offers great potential for our students to experience languages in all their glory.

Resources


#LancsBox corpus toolbox: http://corpora.lancs.ac.uk/lancsbox/

Reverso Context contextual dictionary: https://context.reverso.net/translation/

Sketch Engine, a corpus manager and text analysis tool: https://www.sketchengine.eu/corpora-and-languages/