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

Matt Absalom[1]

Potential impact	medium
Timescale	ongoing
Keywords	digital corpora, corpus linguistics, data-driven learning, language in context

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

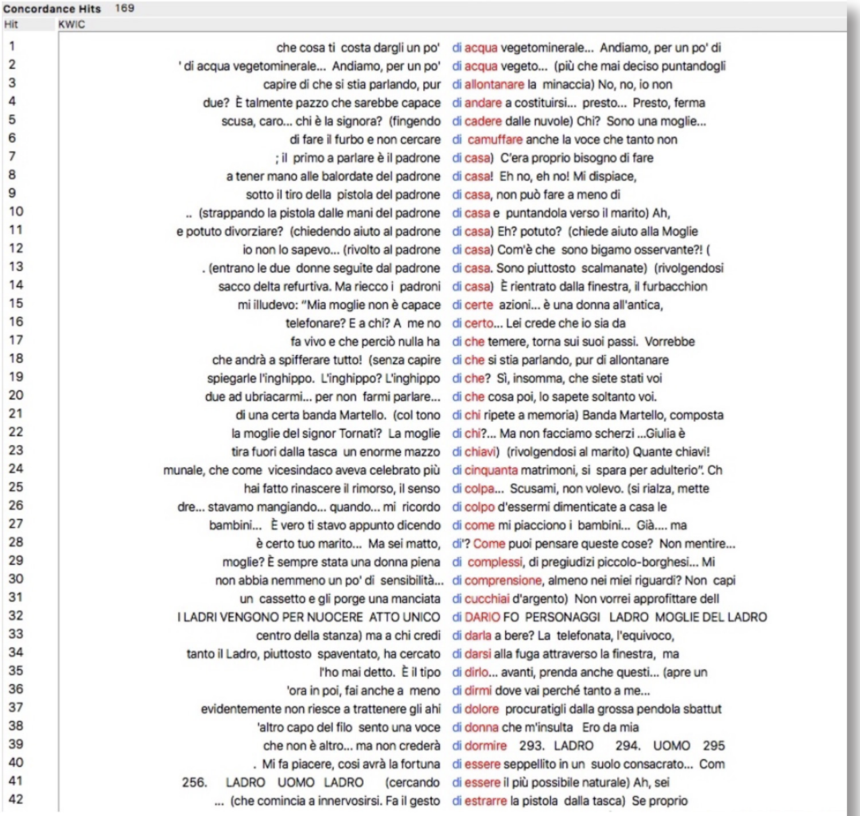
How to cite: Absalom, M. (2021). Digital corpora: language teaching and learning in the age of big data. In T. Beaven & F. Rosell-Aguilar (Eds), *Innovative language pedagogy report* (pp. 97-101). Research-publishing.net. https://doi.org/10.14705/rpnet.2021.50.1242

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

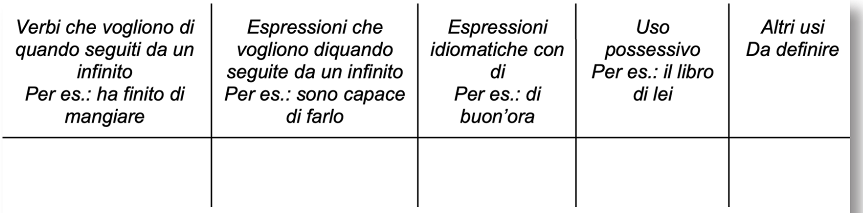
A constant stumbling block for learners of Italian is the choice of preposition. This often comes from the simplistic one-to-one translations presented in language textbooks, manuals, etc. In order to sensitise students to the importance of context in the correct selection of prepositions, I devised 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.



Concordance Hits 169

Hit	KWIC	
1	che cosa ti costa dargli un po'	di acqua vegetominerale... Andiamo, per un po' di
2	' di acqua vegetominerale... Andiamo, per un po'	di acqua vegeto... (più che mai deciso puntandogli
3	capire di che si stia parlando, pur	di allontanare la minaccia) No, no, io non
4	due? È talmente pazzo che sarebbe capace	di andare a costituirsi... presto... Presto, ferma
5	scusa, caro... chi è la signora? (fingendo	di cadere dalle nuvole) Chi? Sono una moglie...
6	di fare il furbo e non cercare	di camuffare anche la voce che tanto non
7	; il primo a parlare è il padrone	di casa) C'era proprio bisogno di fare
8	a tener mano alle balordate del padrone	di casa! Eh no, eh no! Mi dispiace,
9	sotto il tiro della pistola del padrone	di casa, non può fare a meno di
10	.. (strappando la pistola dalle mani del padrone	di casa e puntandola verso il marito) Ah,
11	e potuto divorziare? (chiedendo aiuto al padrone	di casa) Eh? potuto? (chiede aiuto alla Moglie
12	io non lo sapevo... (rivolto al padrone	di casa) Com'è che sono bigamo osservante?! (
13	. (entrano le due donne seguite dal padrone	di casa. Sono piuttosto scalmanate) (rivolgendosi
14	sacco delta refurtiva. Ma riecco i padroni	di casa) È rientrato dalla finestra, il furbacchion
15	mi illudevo: "Mia moglie non è capace	di certe azioni... è una donna all'antica,
16	telefonare? E a chi? A me no	di certo... Lei crede che io sia da
17	fa vivo e che perciò nulla ha	di che temere, torna sui suoi passi. Vorrebbe
18	che andrà a spifferare tutto! (senza capire	di che si stia parlando, pur di allontanare
19	spiegarle l'inghippo. L'inghippo? L'inghippo	di che? Sì, insomma, che siete stati voi
20	due ad ubriacarmi... per non farmi parlare...	di che cosa poi, lo sapete soltanto voi.
21	di una certa banda Martello. (col tono	di chi ripete a memoria) Banda Martello, composta
22	la moglie del signor Tornati? La moglie	di chi?... Ma non facciamo scherzi ...Giulia è
23	tira fuori dalla tasca un enorme mazzo	di chiavi) (rivolgendosi al marito) Quante chiavi!
24	munale, che come vicesindaco aveva celebrato più	di cinquanta matrimoni, si spara per adulterio". Ch
25	hai fatto rinascere il rimorso, il senso	di colpa... Scusami, non volevo. (si rialza, mette
26	dre... stavamo mangiando... quando... mi ricordo	di colpo d'essermi dimenticate a casa le
27	bambini... È vero ti stavo appunto dicendo	di come mi piacciono i bambini... Già.... ma
28	è certo tuo marito... Ma sei matto,	di'? Come puoi pensare queste cose? Non mentire...
29	moglie? È sempre stata una donna piena	di complessi, di pregiudizi piccolo-borghesi... Mi
30	non abbia nemmeno un po' di sensibilità...	di comprensione, almeno nei miei riguardi? Non capi
31	un cassetto e gli porge una manciata	di cucchiai d'argento) Non vorrei approfittare dell
32	I LADRI VENGONO PER NUOCERE ATTO UNICO	di DARIO FO PERSONAGGI LADRO MOGLIE DEL LADRO
33	centro della stanza) ma a chi credi	di darla a bere? La telefonata, l'equivoco,
34	tanto il Ladro, piuttosto spaventato, ha cercato	di darsi alla fuga attraverso la finestra, ma
35	l'ho mai detto. È il tipo	di dirlo... avanti, prenda anche questi... (apre un
36	'ora in poi, fai anche a meno	di dirmi dove vai perché tanto a me...
37	evidentemente non riesce a trattenere gli ahi	di dolore procuratigli dalla grossa pendola sbattut
38	'altro capo del filo sento una voce	di donna che m'insulta Ero da mia
39	che non è altro... ma non crederà	di dormire 293. LADRO 294. UOMO 295
40	. Mi fa piacere, cosi avrà la fortuna	di essere seppellito in un suolo consacrato... Com
41	256. LADRO UOMO LADRO (cercando	di essere il più possibile naturale) Ah, sei
42	... (che comincia a innervosirsi. Fa il gesto	di estrarre la pistola dalla tasca) Se proprio

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 diquando 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

Boulton, A., & Landure, C. (2016). Using corpora in language teaching, learning and use. *Research and Teaching Languages for Specific Purposes, 35*(2). https://doi.org/10.4000/apliut.5433

Flowerdew, J. (2009). Corpora in language teaching. In M. H. Long and C. J. Doughty (Eds), *The handbook of language teaching* (pp. 327-350). Willey Blackwell. https://doi.org/10.1002/9781444315783.ch19

#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/

Research-publishing.net

Published by Research-publishing.net, a not-for-profit association
Contact: info@research-publishing.net

Innovative language pedagogy report
Edited by Tita Beaven and Fernando Rosell-Aguilar

Publication date: 2021/03/22

Disclaimer: Research-publishing.net does not take any responsibility for the content of the pages written by the authors of this book. The authors have recognised that the work described was not published before, or that it was not under consideration for publication elsewhere. While the information in this book is believed to be true and accurate on the date of its going to press, neither the editorial team nor the publisher can accept any legal responsibility for any errors or omissions. The publisher makes no warranty, expressed or implied, with respect to the material contained herein. While Research-publishing.net is committed to publishing works of integrity, the words are the authors' alone.

Trademark notice: product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Copyrighted material: every effort has been made by the editorial team to trace copyright holders and to obtain their permission for the use of copyrighted material in this book. In the event of errors or omissions, please notify the publisher of any corrections that will need to be incorporated in future editions of this book.

Typeset by Research-publishing.net
Cover layout by © 2021 Raphaël Savina (raphael@savina.net)
Photo by Digital Buggu from Pexels (CC0)

ISBN13: 978-2-490057-86-3 (Ebook, PDF, colour)
ISBN13: 978-2-490057-87-0 (Ebook, EPUB, colour)
ISBN13: 978-2-490057-85-6 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: mars 2021.