Compiling and analysing a large corpus of online discussions to explore users’ interactions

Journal Item

How to cite:
Chua, Shi-Min (2022). Compiling and analysing a large corpus of online discussions to explore users’ interactions. Applied Corpus Linguistics, 2(2), article no. 100017.

For guidance on citations see FAQs.

© 2022 Elsevier Ltd. All rights reserved.

https://creativecommons.org/licenses/by/4.0/

Version: Version of Record

Link(s) to article on publisher’s website:
http://dx.doi.org/doi:10.1016/j.acorp.2022.100017

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data policy on reuse of materials please consult the policies page.

oro.open.ac.uk
Compiling and analysing a large corpus of online discussions to explore users’ interactions

Shi Min CHUA

The Open University, Walton Hall, Kents Hill, Milton Keynes, MK7 6AA, England

A R T I C L E   I N F O

Keywords:
Corpus workbench tool
Online discussions
Micro-analysis
Digital conversation analysis
Interactional approach

A B S T R A C T

This methodology-focused paper reports how I compiled and analysed a 12-million-word corpus of threaded online discussions by employing Corpus Workbench tool (CWB, Evert & Hardie, 2011) and combining corpus analysis with micro-analysis drawing on the principles of digital Conversation Analysis. The tool not only affords an efficient retrieval and analysis of a large dataset, but also, more importantly, facilitates exploration of a corpus of online discussions based on different variables (e.g., topics of discussions, role of internet users, types of postings) and units of analysis (e.g., subforums, threads, postings). Examples are presented to illustrate how I used this tool to investigate various aspects of online discussions, and extract threads surrounding a particular topic or language practices for micro-analysis. I propose internet users’ interactions in online discussions can be further explored in the field of corpus linguistics by using this tool and a synergy of corpus linguistics and an interactional approach.

Introduction

Various corpus tools are freely available for researchers to interrogate their corpus, including Antconc (Anthony, 2017), Wordsmith (Scott, 2020), LancsBox (Brezina et al., 2020). With these tools, researchers can easily upload their texts, and within seconds they can start conducting keyword analysis, collocation analysis and concordance reading. However, such tools may not be suitable for handling large corpora of more than 1 million words. More importantly, these tools might be cumbersome for compiling and analysing data which can be explored based on various variables and units of analysis. One such data is online discussions which consist of multiple conversation threads participated in by many users and are organized under numerous topics in a single forum or site. Various research questions can be asked at the different levels of online discussions to provide insights into users’ language practices and their interactions with each other. However, previous corpus analysis typically treated all comments in an online discussion as the same without any differentiation, for example differentiation between replies and new posts, posts that receive replies vs. those that do not, or posts that are contributed by different groups of users (Drasovean and Tagg, 2015; Sotillo and Wang-Gempp, 2016).

In this paper, I show how I employed Corpus Workbench tool (CWB, Evert and Hardie, 2011) to compile and explore language practices of users’ interactions in a 12-million-word corpus of threaded online discussions in massive open online courses (MOOCs). This approach allows comparison of different types of posts that lead to revelation of various online language practices, including those that are more likely to trigger responses from other users, as well as language practices employed by different groups of users and in different situations. Additionally, I demonstrate how this tool can facilitate the selection of threads from a large corpus for a more in-depth turn-by-turn analysis, in line with the methodology of micro-analysis and digital Conversation Analysis (digital CA, Giles et al., 2014). Lastly, I argue for complementing corpus analysis with an interactional approach, such as micro-analysis or digital CA, to better understand language practices in online interactions, similar to O’Keeffe & Walsh (2012) who combined corpus analysis and conversation analysis to analyze interaction patterns and language use in face-to-face classroom discussions. This integrated approach enriches linguistic analyses of online discussions by expanding the possibility of keyword analysis to comparisons based on variables relevant to online interactions and by deepening the investigation to a turn-by-turn analysis that reveals the impact of language practices on the on-going online conversations.

Corpus analysis of online discussions

Since the inception of internet, online discussions have provided a tremendous amount of textual data for researchers to explore language and communicative practices in this new medium, as well as internet users’ opinions and discursive construction of social issues (Herring, 1999). Online discussions are not only found in traditional online spaces dedicated for discussions, such as forums and sup-
port/interest groups, newsgroups, bulletin board systems, mailing lists and social networking sites, but also in the commenting space alongside multimedia content such as in blogs, news-websites, TED, Facebook or YouTube.

Many researchers have utilized corpus linguistics methodology that affords both quantitative analysis and qualitative analysis to investigate the big textual data in online discussions (Drasovean and Tagg, 2015; Hunt and Harvey, 2015; McDonald and Woodward-Kron, 2016; Paterson, 2020; Sotillo and Wang-Gempp, 2016). For example, by investigating collocates of the keywords referring to the main topics in the respective discussions, that is anorexia and politicians, Hunt & Harvey (2015) and Sotillo & Wang-Gempp (2016) provided insights into users’ experience and conceptualization of the topics, while Drasovean & Tagg (2015) revealed language practices that establish online communities.

In these corpus analyses of online discussions, the keyword analysis and collocation analysis are typically conducted on all the comments in the discussions while the concordance reading is limited to individual comments instead of threads. This analysis strategy focuses mainly on the language use in the online discussions as a whole, while the interactive aspects of online discussions have not yet been examined, for example, how users respond to each other, sustained conversation among users on a particular topic, different types of postings (e.g., posts that receive replies and those that do not, replies) and language practices of users with different roles (e.g., moderators, prolific users). Although Drasovean & Tagg (2015) and Sotillo & Wang-Gempp (2016) revealed various discourse practices for challenging or entertaining others’ voices in online discussions, there was no mention of the responses or the messages being responded to; that is, which comments are challenged or entertained. Only two studies have compared different types of postings in online discussions. McDonald & Woodward-Kron (2016) examined language practices of new and veteran members in an online support group. They divided their corpus of online discussions into subcorpora of postings based on individual users’ post count at the time of posting. Subcorpus 1 consists of all users’ first posts, subcorpus 2 consists of all users’ second and third posts and so on. Collins (2019) compared the posts and replies in an online discussion of MOOCs, but did not go into details to explore discourse practices realized by the keywords found. Both studies only examined the language practices in individual comments, rather than the interactions, that is how the new and veteran members respond to each other or the relationships between posts and replies.

2.1. Challenges of corpus analysis of online discussions

The interactive aspects of online discussions, coupled with the threaded discussions and potentially large number of postings could present particular technical challenges to researchers interested in conducting corpus analysis of online discussions (Claridge, 2006). Typically, there are multiple discussion threads under different forum topics. For example, within a finance forum, there might be various related topics, such as pension and housing. Within each topic, there are several threads, which are initiated by new posts that receive replies, and some threads may drift away from the original topics. There are also posts that do not receive any replies, thus not generating a thread. Various research questions can be asked at the different levels of the online discussions to further our understanding of online interactions. However, because of the unique structure of online discussions, exploration of different variables (e.g., topics of discussions, role of internet users, types of postings) and units of analysis (e.g., subforums, threads, individual postings, words) may not be easily conducted with currently available corpus tools. For example, to conduct keyword analysis of new posts and replies, researchers may need to save all the new posts into one text file as one subcorpus, and replies into another text file as another subcorpus, rather than keeping them in the same corpus to maintain the original threads. Similarly, to investigate language practices of moderators and users, researchers may need to save the discussion postings of different groups of users into separate text files as different subcorpora. Therefore, exploration of multiple aspects of online discussions could become tedious because investigation of each variable requires a new organization of corpus files and subcorpora.

Furthermore, because postings from the same threads are saved into different subcorpora during the keyword analysis comparing different types of postings, concordance reading might be limited to the specific postings rather than the whole threads, unless researchers relate the postings back to the threads. At times, even if the whole threads are maintained, the corpus tools might not allow encoding of their structure or extension of concordance lines to the defined larger context, such as the threads or the postings before and after the particular postings investigated. Yet, researchers can better understand the language practices adopted in each posting by taking into account the context, i.e., the thread where a particular posting occurs. Threads are where explicit user-user interactions and dialog occurs; that is, where users respond to each other, therefore it might be best to examine the threads to understand online language-in-action (Ksiazek and Lessard, 2016). With respect to keywords in replies, concordance readings that are limited to the replies are decontextualized in the sense that replies are always in response to postings posted before them and may trigger other replies posted after. Therefore, the communicative functions of the keywords within a reply might be better interpreted in light of the whole thread, and preferably with an interactional approach.

In short, extending concordance lines across different units of analysis, that is from lines to postings to whole threads, is an important step in the corpus analysis of online discussions. Yet, most corpus tools may not allow such flexibility and the necessary annotation for considering the different units of analysis. In the following, I will first introduce the principle of an interactional approach to online discussions, then demonstrate how CWB can be used to resolve these challenges of conducting corpus analysis on data from online discussions and to incorporate an interactional approach in the analysis.

2.2. Interactional approach to online discussions

One interactional approach to online discussions is digital CA, a methodology developed by Giles et al. (2014) who adapted the principles of (spoken) conversation analysis (CA) to examine online data. Essentially, this approach analyses a conversation turn-by-turn to understand social interactions (Sacks et al., 1974). A conversation can be seen as a sequence of actions or conversational moves, such that each turn can be considered as an immediate response to the preceding turn, and at the same time be designed to elicit response in the next turn. Thus, CA analyses the language practice in each turn in relation to the adjacent turn. In digital CA, the post and replies within a thread can be loosely operationalized as corresponding to the turn-taking in spoken conversation, although users do not need to compete for conversational floor, thus turn-taking within a thread can interleave with multiple conversations and a turn can be one-to-many (Herring, 1999; König, 2019).

Unlike corpus analysis which exploits big data, digital CA is a micro-level analysis that examines a thread in depth. For example, Stommel & Koole (2010) conducted digital CA on a thread of a forum in which a new member interacts with other members. They showed that mismatch of turn-taking between new members and veteran members may result in new members not being welcomed. This happens when the new member engages in problem telling in their opening post (e.g., “I had to tell someone”, p. 366) but does not acknowledge the advice given by veteran members who treat the opening post as advice seeking. This micro-level analysis provides insights into the implications of a particular language practice on the interactions between new and veteran members. This is a perspective different from McDonald & Woodward-Kron’s (2016) corpus analysis that only shows the differences in language choices adopted by new and veteran members. This micro-level analysis highlights the importance of examining the whole thread of online discussions on top
of the corpus analysis, that is expanding the concordance lines to the threads.

Micro-analysis is another interactional approach widely used to examine discourse practices in online discussions. The analysis adopts digital CA’s focus on how users respond to each other and does not focus solely on adjacent pairs. For example, Ziegler et al. (2014) examined an online discussion thread through micro-analysis and revealed discourse practices employed by users to explore new ideas and engage in negotiation, such as making underlying assumption explicit and raising new questions.

The potential of combining corpus linguistics with digital CA or micro-analysis can also be gleaned from O’Keeffe & Walsh’s (2012) approach that combines corpus linguistics with conversation analysis (CLCA) to analyze face-to-face classroom interactions. Their CLCA approach lies in paying attention to the keywords found by corpus analysis while conducting CA of episodes of classroom interactions. This allows them to better understand why certain keywords are used in certain turns in classroom interactions and gain more insight into the communicative function of the keywords. For example, they found that one of the keywords in their corpus, okay, is often used in the turns where the tutors inform the students about the organizations of the modules and upcoming activities.

These previous studies show that both micro-analysis and digital CA of online discussion are useful in revealing discourse practices among users in their interactions. Therefore, conducting either analysis on top of corpus analysis might reveal further the interactional functions of the keywords found. In short, to effectively conduct a corpus analysis of online discussions with an interactional approach, a corpus tool or procedure that allows exploration of various aspects of online interactions, including different variables (e.g., topics of discussions, role of internet users, types of postings) and units of analysis (e.g., subforums, threads, individual postings, words) is needed. In the following, I show how I utilized CWB to compile and explore a 12-million-word corpus of threaded online discussions. I elaborate on how the tool can be used to facilitate a synergy of corpus analysis and interactional approach in addition to what has been proposed by O’Keeffe & Walsh’s (2012) CLCA approach.

The corpus of online discussions in this study – MOOC corpus

The data for the MOOC corpus comes from the online discussions on a massive open online course (MOOCs) platform – FutureLearn. The MOOCs typically run for two to eight weeks. Each week consists of several learning pages. Each learning page contains one main learning object, which can be a video, audio, article or discussion prompt. Below the learning content, there is a commenting space that allows users to post their comments, similar to those underneath YouTube videos or news website articles. Parallel to other online discussions, each MOOC can be considered a forum, while each learning page is a subforum. Learners contribute much more than facilitators, who are usually the academics designing the course. The corpus comprises 221,823 postings with a total wordcount of 12,022,278 in the commenting space from 12 MOOCs. The number of comments and words contributed by learners and facilitators across the 12 MOOCs can be found in Table 1.

In the commenting space, a user can create a new post or reply to others’ posts or replies. I use the term comments to refer to posts and replies collectively. Each new post has the potential to elicit replies from other users and initiate a thread of discussion, which is where explicit user-user conversation occurs (Książecki and Lessard, 2016). I call the posts that receive replies initiating posts (n = 32,333). However, most posts are left unresponded, which I call independent posts (n = 118,606). The number of initiating posts, independent posts and replies, and corresponding word count contributed by learners and facilitators across the 12 MOOCs can be found in Table 2.

Most threads have fewer than five replies (n = 29,354), although there are threads containing more than 20 replies (n = 41), as shown in Fig. 1. There is no threaded structure among replies underneath a post,
Table 2

<table>
<thead>
<tr>
<th>Facilitator</th>
<th>MOOCs</th>
<th>Independent post</th>
<th>Number of Comments</th>
<th>Word count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>accessibility-2</td>
<td>1151</td>
<td>92880</td>
<td>4731</td>
</tr>
<tr>
<td></td>
<td>arctic-cent-1</td>
<td>1010</td>
<td>129089</td>
<td>1470</td>
</tr>
<tr>
<td></td>
<td>carrot-1</td>
<td>946</td>
<td>113649</td>
<td>1510</td>
</tr>
<tr>
<td></td>
<td>denim-1</td>
<td>946</td>
<td>106987</td>
<td>1290</td>
</tr>
<tr>
<td></td>
<td>fin-1</td>
<td>1845</td>
<td>188831</td>
<td>7349</td>
</tr>
<tr>
<td></td>
<td>management-1</td>
<td>1207</td>
<td>17349</td>
<td>7349</td>
</tr>
<tr>
<td></td>
<td>moss-1</td>
<td>4386</td>
<td>120854</td>
<td>11734</td>
</tr>
<tr>
<td></td>
<td>ocean-1</td>
<td>1626</td>
<td>19117</td>
<td>2211</td>
</tr>
<tr>
<td></td>
<td>palmer-1</td>
<td>720</td>
<td>92117</td>
<td>12917</td>
</tr>
<tr>
<td></td>
<td>tea-1</td>
<td>801</td>
<td>101813</td>
<td>13717</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2205</td>
<td>2409208</td>
<td>2409208</td>
</tr>
</tbody>
</table>

Table 3

<table>
<thead>
<tr>
<th>Units of Analysis</th>
<th>Potential Variables (for keyword analysis)</th>
<th>Suggested focus for analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOOCs</td>
<td>• Humanities vs. Sciences MOOCs</td>
<td>• Language practices in the discussions of different disciplines</td>
</tr>
<tr>
<td>WEEKS</td>
<td>• Week 1, 2, 3, ...</td>
<td>• The change of language practices and vocabulary usage as users progress through the course</td>
</tr>
<tr>
<td>Learning</td>
<td>• First week vs. final week</td>
<td>• The discussions triggered by different learning objects</td>
</tr>
<tr>
<td>Pages</td>
<td>• Learning page with video vs. article</td>
<td>• Language practices or topics that generate sustained conversations among users</td>
</tr>
<tr>
<td>Threads</td>
<td>• Long threads vs. short threads</td>
<td>• The impact of facilitators on conversations</td>
</tr>
<tr>
<td>Comments</td>
<td>• Initiating posts vs. independent posts</td>
<td>• Language practices in different types of comments</td>
</tr>
<tr>
<td></td>
<td>vs. replies</td>
<td>• Language practices of different groups of users</td>
</tr>
</tbody>
</table>

and the replies are ordered by the time of posting. The interface of the online discussions is shown in Fig. 2.

Potential units of analysis

Five units of analysis can be derived from this online discussion, as summarized in Table 3, along with suggestions for corpus analysis. For example, a researcher can conduct a keyword analysis at the unit of MOOC comparing a Humanities MOOC with a Science MOOC to examine language practices in the online discussions of different disciplines. A keyword analysis at the unit of learning page can compare discussion postings underneath a learning page with video to a page with a discussion prompt to reveal whether different learning objects generate interactions of different nature among users, as indicated by their language practices. In my study, two keyword analyses were conducted at the unit of comments. First, a keyword analysis between initiating posts and independent posts revealed language practices that are more likely to trigger responses in online discussions (Chua, 2021). Second, a keyword analysis of facilitators’ postings revealed how facilitators affect the online discussions (Chua, 2018).

Employing Corpus workbench tool (CWB)

The data, i.e., the text of the comments, and metadata, i.e., the various variables, were encoded into CWB (https://cbw.sourceforge.io/index.php; Evert and Hardie, 2011). The CWB is a corpus tool with efficient data structure and retrieval for concordancing as well as frequency counts and sorting. It is the query architecture underlying the CQP web (Hardie, 2012) that hosts various corpora for the public and the commercial corpus tool, SketchEngine (Kilgarriff et al., 2014). Although CWB might be less user-friendly than other corpus tools such as Antconc (Anthony, 2017), Wordsmith (Scott, 2020) and LanesBox (Brezina et al., 2020), it is more efficient when it comes to big data. Most tools require data to be loaded every time they open, and are time-consuming and unstable for big data, whereas CWB only requires the data to be encoded once and can manage large corpora ranging from 10 million to 2 billion words. Therefore, it is more suitable for a large corpus, such as the present corpus of online discussions.

More importantly, CWB allows a data structure that can authentically encode the online discussions. As mentioned earlier, the comments in the online discussions are nested within five units of analysis: MOOCs, Weeks, Learning Pages, Threads, Comments. They were annotated in XML according to TEI standard (TEI Consortium, 2020), as shown in Fig. 3, then subjected to treetagger (Schmid, 1994) for tokenization, lemmatization and part-of-speech tagging before being encoded in the
Fig. 1. Number of threads of different lengths across the 12 MOOCs. Each tile indicates a set of threads of a particular length. The shading of the tiles indicates the number of threads that correspond with that particular length in a particular MOOC. The length of threads starts from 2 as the shortest thread contain one initiating post and one reply.

Table 4
Data provided by the MOOC providers

<table>
<thead>
<tr>
<th>Data</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>The id of the comment.</td>
</tr>
<tr>
<td>author_id</td>
<td>The user’s id.</td>
</tr>
<tr>
<td>parent_id</td>
<td>The id of the post that the comment posted under. If there is a parent_id, the comment is a reply, if it is blank, it is either an initiating post or independent post.</td>
</tr>
<tr>
<td>step</td>
<td>The step where the comment is posted.</td>
</tr>
<tr>
<td>week_number</td>
<td>The week of the course where the comment is posted.</td>
</tr>
<tr>
<td>step_number</td>
<td>The learning page in the week where the comment is posted.</td>
</tr>
<tr>
<td>text</td>
<td>The comment.</td>
</tr>
<tr>
<td>timestamp</td>
<td>The date and time when the comment is posted.</td>
</tr>
<tr>
<td>likes</td>
<td>The number of likes the comment receives.</td>
</tr>
</tbody>
</table>

CWB. More details about the encoding procedures can be found in the CWB manual (Evert, 2021).

Annotation

Typically, the annotation of meta-data can be achieved via two methods: human judgement and automation, depending on the nature and availability of the meta-data. For example, whether a learning page mainly consists of videos, articles, audios, or discussion prompts may require researchers’ judgment because such data is not provided on the platform. In contrast, the annotation of the meta-data for the five units proposed above can be automatized because most of them can be extracted or calculated based on the data files provided by the MOOC providers (Table 4). For other online discussions, it may well depend on the data provided via the application programming interface (API) or scraped directly from the webpages.

As shown in Table 4, the data provided did not include information regarding the types of the comments, i.e., initiating posts, independent posts, replies, and whether they were contributed by facilitators. Based on the id, parent_id, author_id and timestamp, I was able to categorize the comments into the three types of comments, i.e., initiating posts, independent posts, and the replies, and arrange them within the threads according to the time of posting. This categorization was automatized by programming software such as R. The calculation of the length of the thread was also achieved by R. Only the identification of facilitators required some manual work. I visited each MOOC on the platform and checked the introductory page where facilitators were introduced, and browsed through the discussion space in the first week of the courses to gather their comments. I then mapped them back to the data file to identify their id accordingly, which is then used in R to create a variable to annotate whether a comment is contributed by a learner or facilitator, and whether a thread involves facilitators. Lastly, I used R to export the data into XML (Fig. 2 presented earlier).

Conducting corpus analysis based on different variables

All the annotated and corresponding variables (e.g., initiating posts vs. independent posts) can then be used in queries in the tool with the command line \texttt{[/:match.(variables) where \texttt{[}/ refers to all words or can be any words of interest by typing \texttt{[word=2X]%cd} while more than one variable can be matched. The concordance lines of the query will then be shown and could be sorted (Fig. 4). For example, the command line \texttt{[/:match.comment_type=“initiating post” searches all the words of initiating posts. Following it with \texttt{count Last by word%cd}, the word frequency of all the words can be tabulated (Fig. 5). Furthermore, the distribution of word frequency across another variable (Fig. 6), e.g., MOOC of different subjects, can easily be shown with the command line \texttt{group Last match (variables)}, e.g., \texttt{group Last match mooc_id}. In the tool, \texttt{Last} refers to the search conducted. More functions of CWB, including query for collocates, can be found in its manual (Evert, 2021).

The feature of matching based on a variable or unit of analysis during the query is essential for the various keyword analyses based on variables mentioned above. After the query, the feature of grouping can
Over the last three weeks you have learned that inequalities in access to housing and in access to a decent pension in old age are influenced by and have an impact on wider socio-economic inequalities. This week returns to these wider inequalities and explores the relationships between different levels of intervention, and how these have evolved over time and could evolve in the future.

Kick-start this week by reminding yourself what is at stake.

© The Open University

118 comments

Fig. 2. The commenting space below each learning page of a MOOC.

further show the distribution of word(s) across another variable, thus facilitating examination of the word(s) based on multiple variables. The tool does not require researchers to save comments separately based on different levels of a variable to make subcorpora for comparison, thus reducing the tedious menial tasks while encouraging exploration of data. However, the CWB does not offer statistical analysis such as log-likelihood ratio test for keyword analysis and mutual information for collocation analysis, as provided by other corpus tools. Therefore, the output of the frequency counts will have to be subjected to other programs, such as R for statistical analysis.
By employing CWB and R, I explored my corpus by conducting several keyword analyses based on variables related to users’ interactions. Firstly, in one keyword analysis between initiating posts and independent posts (Chua, 2021), I found differences between these two types of posts, and identified language practices that are more likely to trigger replies based on the keywords of initiating posts. One of the practices is using modals (might, would, could), hedges (perhaps, seems, sort of) or expressions of uncertainty (I wonder, I am wondering), not knowing (I don’t know), or possible mistakes (I might be wrong, am I missing something?) to soften one’s claim and avoid bare assertions. The tentativeness expressed by these keywords acknowledges the possibility of alternative voices and creates spaces for others to pitch in, thereby increasing the
chance of receiving replies from others (Martin and White, 2005). In contrast, when I compared the keywords in replies between those in long threads (five or more than five replies) and those in short threads (fewer than five replies), no difference was found, suggesting either language does not play a role in sustaining a conversation or language practices sustaining a thread are not realized by particular keywords or expressions.

**Expanding concordances to different units of analysis**

The tool allows expanding of the concordance lines based on the units of analysis encoded, with the command line set Context (unit of analysis) before typing the search query. For example, to expand the concordance to the whole comment or thread, the command set Context comment or thread can be used (Fig. 7). This way, researchers can easily examine the wider context where a word is used, instead of just a few words around it. Alternatively, researchers could still use the typical concordance lines without expanding the context but choose to show the id of the thread or comment in which a word/keyphrase is found by using the command line set PrintStructures thread_id or comment_id. Another way is to use the command line group Last match thread_id or comment_id to gather the id of all the threads or comments (Fig. 8). This way, researchers can easily refer to their corpus for the whole threads or comments. For example, I used the id to extract threads or comments into individual files to be subjected to Nvivo program that allows me to make notes of my analysis on the text. Expanding context in which a word/keyphrase appears not only facilitate the analysis of the communicative functions of the words but also allows researchers to conduct a CLCA-like analysis suggested by O’Keefe & Walsh (2012). By using the above-mentioned CWB functions, the researchers can easily collect all the threads in which a keyword appears and conduct micro-analysis or digital CA.

**Case studies: combining corpus analysis with micro-analysis**

In this section, I present two ways that online discussions can be explored with a synergy of corpus analysis and interactional approach by using the CWB tool. The first analysis reveals the practice of concessions in online discussions by conducting a micro-analysis on a thread that contains keywords identified from a keyword analysis. The second analysis unravels in-depth discussions on a specific topic among users by extracting threads with the most mentions of specific terms.

**Interpreting keywords for their communicative functions within a thread**

In the first analysis, the combination of corpus analysis with micro-analysis starts from a keyword analysis of replies (by comparing replies to initiating posts and independent posts, see Chua, 2021 for details), then moves on to an in-depth analysis of a thread by paying attention to keywords found, similar to the CLCA approach to face-to-face interactions (O’Keefe and Walsh, 2012). That is, although I have a glimpse of the language practices in replies based on the keywords found, I move on to conduct a micro-analysis of threads to understand the implication of the language practices realized by the multiple keywords in online interactions among users. For example, among the 57 keywords found in replies, based on my concordance reading, the following are typically used to align with what other users have said: boosters (exactly, absolutely, totally, indeed), evaluatives (true, right), attitude verbs (agree, agreed), especially in the form of agree with you/your..., and polite speech-act formulae (yes). However, an in-depth analysis of the threads where these keywords are used together with another keyword but, shows that they are used for concession. Specifically in the turn-taking between users who hold different points of views, users use these keywords together to concede what others have said before tabling their own views. Furthermore, a micro-analysis shows that concession strategies facilitate users’ negotiation with each other in online discussions, as we shall see in the thread presented in excerpt 1 below. The original wordings by users, including misspellings and typos, are maintained in the excerpts while the replies keywords are italicized.
This thread was first identified when I queried for all keywords in replies and used the command group Last match thread id to tabulate the frequency of keywords found in each thread. The thread was also one of the ten longest threads in the contract-4 course. It was selected based on the assumption that long threads are more likely to contain turn-taking by users who continue to reply to each other.

Excerpt 1
Initiating post
User m4–285
(First part of the post is omitted as it is not relevant for the current analysis)

I had no idea that a «smell» could be a trademark... (sound I am guessing like some of the iphone rings or notification sounds for example... But smell? Does anyone have an example? And what about taste then?

Reply 1
User m4–285
Ohhh! A perfume!! Like Chanel #5 ... Ok nevermind that! So what about taste?

Reply 2
User m4–153
Cocacola?

Reply 3
User m4–285
Mmmmm Ok... But then who decides, in these cases, that the taste is too similar to Coca Cola of that this fake perfume is too similar to the original one... There is generic Cola... There are perfume copies... I mean, when it comes to smell and taste isn’t it partly subjective? I have hear before cases where there is litigation for a song that is said to have been copied and that violated copyrights and then they send an expert on harmonics or something to evaluate the level of similarity with the «originals»... But taste and smell... ??

Reply 4
User m4–153
Yes, there are copies, but they cannot be the same as the real thing. A very good imitation LV handbags, no matter how good it was copied can never be the same as the real thing. Each LV bag is hand made and the monograms are arranged in a unique way. even though the imitated ones are made of quality leathers, such little details give it away. Also for taste, there are thousands of other cola drinks, but they are never the same. Channel No 5 might be imitated, but it wouldn’t smell the same as. also the longevity of the smell might differ too.

Reply 5
User m4–285
That is true but that is not exactly what I was wondering... Is taste really a trademark if you can replicate it? Coca cola does not go behind generic colas and sue them for violating their property. Nor does the perfume companies... Do they have people specialized in evaluating the other colas and say: oh ok, it tastes different enough from Coca Cola...? Same for perfume...

Reply 6
User m4–153
I understand your stance. They cant be sued because they don’t taste or smell the same. Close enough, but not the same.

Reply 7
User m4–285
Who draws that line?

Reply 8
User m4–153
The Patent, TM or IP details does. I suppose.

Reply 9
User m4–443
I was wondering about this - must be very difficult to draw an agreed line in cases where there is an area of subjectivity.

Reply 10
User m4–153
True. But Subjectivity is Individually Contraled

Reply 11
User m4–285
Exactly!

This thread contains 11 replies involving mainly two users who repeatedly come back to the same thread, and there are 35 tokens of keywords in the replies. Two keywords but (n = 8) and same (n = 7) are used frequently by both users in their discussion. But is always used in a concession practice in their negotiation with each other where they re-instate their own view after acknowledging the other’s view with attitude keywords such as oh, yes and true. Exactly and too are also used by users to emphasize their own stance. Same is used in relation to the topic of discussion, whether a fake product is the same as the original product and the frequent usage is likely to be specific to this thread only.

The language practice of conceding and reasserting by using the keywords expressing agreement and but can be observed in reply 3 to reply 6. In their responses to each other, user m4–285 and user m4–153 concede by first agreeing with the other’s viewpoint, then reassert their own view with the keyword but, and challenge the other’s view. The concession in reply 3 is indicated by “Mmmmm Ok” in which Ok could mean agreement, yet the following “But then who decides, in these cases, that the taste is too similar to Coca Cola” indicates user m4–285’s reassertion and disagreement with m4–153’s response in reply 2 “Cocacola?”.
Following this concession and reassertion, the user m4–285 also further elaborates on their stance by raising a rhetorical question “when it comes to smell and taste isn’t it partly subjective?”, and presenting the contrasting example of song. It could therefore be argued that “Mmmm Ok but...” is a concession strategy that is used to engage with the previous utterance and other readers before expressing disagreement and reasserting one’s own stance. Similarly, in response, user m4–153 also makes a concession “Yes...but...” in reply 4 before launching into a series of examples to reassert their claim that copies are “never the same”. Again, in response, user m4–285 first uses a concession to engage with user m4–153’s response “That is true but...” in reply 5 before repeating their initial question in the initiating post but with a condition to refine it to be more specific, “is taste really a trademark if you can replicate it?”. In turn in reply 6, user m4–153 responds with another concession strategy, “I understand your stance” yet again restates their own claim with but, not and n’t.

This “yes...but” practice is similar to the other-trigger concession found in oral conversations (Lindström and Londen, 2013). Other-trigger concession is a response that acknowledges other’s opposing views, thus suggesting the dialogic nature of the concession practice in the online discussion, despite its asynchronous nature. Through concessions, others’ voices are also referred to in one’s replies, thus ensuring the coherence of the discussion thread while users pitch in more content. More importantly, this concession strategy facilitates the expression of disagreement by taking into account each other’s opinions. Prefacing one’s reply with agreement maintains one’s interpersonal relationship with others, given that bald disagreement can be face-threatening, especially in online communications that lacks nonverbal cues (Baym, 1996). All these point to the importance of concession in online discussions when disagreement arises, as it allows both users to elaborate on their stance in relation to the other’s view and creates a constructive and interactive discussion space. As shown towards the end of this discussion thread, users seem to reach mutual understanding after several rounds of concession and reassertion.

Methodologically, this example shows that integrating corpus analysis with micro-analysis goes one step further than keyword analysis to provide insights into users’ language practices through which they co-construct their interactions. Importantly, it shows how different keywords are combined to realize a language practice, how the same language practices are employed repeatedly within one thread, and how the language practices impact users’ interactions as their discussion evolves, in this case as two users with opposing views come to an agreement. This observation would not be possible if each keyword was only analysed independently and if only restricted context is examined during concordance reading. The micro-analysis enriches the findings of the keyword analysis as it further illustrates the language practices realized by the keywords and their role in users’ interactions. The combination of corpus analysis and micro-analysis is made easy with the CWB tool that allows the extending of concordance lines and the grouping of findings based on the unit of analysis and variable (e.g., replies) at the thread level.

Building collections of threads for micro-analysis or digital CA

The second analysis focuses on the interactional approach and uses CWB to facilitate the extraction of candidate threads from the large corpus of online discussions. To conduct digital CA, researchers follow a systematic procedure to ensure their findings reveal what can possibly be achieved by different language practices, similar to that of CA. In CA, according to Heritage (2004), researchers can first identify some candidate practices or topics which are distinct and relevant to their research interest, by preliminary reading of the data. Then, a collection of conversations that involve these practices are compiled. To analyze the data following CA principles, relevant turns are located in each conversation and analysed. Finally, the researchers may narrow down their choices of candidate practices to the practices they systematically find in the collection of conversations. Similar procedures are used in micro-analysis except that it does not strictly follow the CA principles but maintains the underlying principles of an interactional approach by examining how users respond to each other in a thread. It is in the building of collections of threads that the CWB tool becomes handy, especially when the corpus is so huge that it is not possible for researchers to read through all the threads to decide which threads to be examined closely.

The extraction of candidate threads can be done by using any corpus tools with appropriate search terms. Yet, most tools would likely not reveal the number of times the search terms appear within a thread, which in contrast is afforded by the CWB tool’s grouping function. The number of times the search terms appear within a thread becomes an important point of entry when there are too many threads containing the search terms. Researchers could start from the threads with the most mentions of the search terms based on the assumption that topics related to the search terms will be salient in these threads, then take a saturation approach by continuing to read the threads with decreasing numbers of mentions until no new pattern emerges. For example, in my investigations of URL-posting practices in online discussions (Chua, 2021), there were 3724 threads and 3429 independent posts containing at least one mention of the search term – URL address and link(s). Therefore, although corpus tools help extracting candidate threads for digital CA or micro-analysis, the number can still remain large and requires further pruning before researchers start analysing the threads in-depth. In this case, I started from threads containing the highest number of the mentions of these two search terms, and continued reading the threads with decreasing number of mentions to examine URL-posting practices.

The thread in excerpt 2 illustrates one of the findings from my micro-analysis of URL-posting – link war, a term I create to describe interaction pattern in which disagreeing users exchange URLs that in turn hinders their negotiation process (Chua, 2021). This thread is from a MOOC on nutrition and contains ten URL addresses posted by three disagreeing users who hold different views towards fat and health, two hold strong views for and against and one is on middle ground.

Excerpt 2

Initiating post
User n4–2511
I only use coconut oil to cook with. I take a tablespoon daily (when I remember). I avoid margarine and butters. [other text omitted]
[Reply 1 is omitted]
Reply 2
User n4–211
Eat it. Much better for you than seed oils and polyunsaturates..https://www.ncbi.nlm.nih.gov/pubmed/24723079
Reply 3
User n4–1657
Doctors have been reverting (yes REVERTING) diabetes-2 and heart disease through avoiding all high fat sources. See e.g. Fig. 2, in http://dresselstyn.com/JFP_06307_Article1.pdf See http://drmdougall.com/ or e.g. http://pcrm.org/ From this it would appear we are really a low fat species.

Reply 4
User n4–2611
[n4–211], if you’re referring to coconut oil, the British Nutrition Association, in October 2016 recommended the low consumption of coconut oil as it’s high in saturated fats and so far there’s no evidence of its health benefits

Like with everything, I think moderation is best.
Reply 5
User n4–211
There is no problem with saturated [sic] fats. This is part of the problem. There is no conclusive evidence saturated fat is harmful. This
is currently being challenged to PHE and other nutritional advisories as their recommendations are not backed by research. Heart health is not affected by saturated fat either. Most of this comes from poor science initially done in 1970s by Ansel Keyes. see http://articles.mercola.com/sites/articles/archive/2016/06/05/saturated-fat-heart-disease-risk.aspx

Reply 6
User n4–2511

Many health care advisors advocate eating virgin/organic coconut oil for its health benefits even though it is high in saturated fats. Coconut oil contains lauric acid, which is a medium-chain fatty acid, that converts to monolaurin. Monolaurin is the compound found in breast milk that strengthens a baby’s immunity, and a great deal of research has been done to establish the ability of lauric acid to enhance immunity. This medium-chain fatty acid (MCFA) actually disrupts the lipid membranes of offending organisms such as yeast, fungal and bacteria living in our gut. This is the main reason why I consume it.

Reply 7
User n4–2611

I know about the poor evidence in saturated fat but as I understand there’s not enough research (as far as I know) done on coconut oil to prove its benefits.

Reply 8
User n4–2611

Also, I have found a very interesting article discussing the systematic reviews carried out on research in this field which I think it’s worth looking at. http://www.cebm.net/evidence-really-not-support-introduction-low-fat-dietary-guidance-1983/

It may well be the case that saturated fat is not that bad for you but for now I’ll take the WHO, Public Health England and the British Heart Foundation’s advice :) 

Reply 9
User n4–1657

"Like with everything, I think moderation is best."

Please realize that we will only know the true meaning of moderation if we know the extremes: the healthiest food and the unhealthiest food. Without knowing the extremes ‘moderation’ means putting your head in the sand.

"There is no problem with saturated [sic] fats."
Sorry, if we can REVERT both heart disease and diabetes-2 through a truly low fat diet than there is a definite problem with fats, including saturated fats.

Reply 10
User n4–1657

"There is no conclusive evidence saturated fat is harmful’ Are you sure?? Where do you get that from?? If you can revert heart disease AND diabetes-2 through a truly low fat diet (including very low saturated fats) than I would think that strongly indicates that fats are not just bad but really bad. I wonder if the critics of Ansel Keys really understand a truly low fat diet.

Reply 11
User n4–1657

Here is a good presentation on fat and health research:
https://www.youtube.com/watch?v=Lbtww2P4Yfs

Reply 12
User n4–211

Some of the quoted studies here are from 2000 and old research. There was the presence of high carb too which have more recently been indicated to cause very low density lipoprotein which are the “bad” part of LDL. Also these were funded by big pharma in view of supprting [sic] statin sales. Sorry not convinced.

Reply 13
User n4–1657

The newer the research, the bigger the commercial influence. If "Doctors have been reverting (yes REVERTING) diabetes-2 and heart disease through avoiding all high fat sources. See e.g Fig. 2, in http://dresselstyn.com/JFP_06307_Article1.pdf See http://drmcclougall.com/ or e.g. http://pcrm.org/ From this it would appear we are really a low fat species.” does not convince you than nothing will.

The micro-analysis of the thread illustrates two characteristics of link war. Firstly, users employ URLs as their main argument for discussions. This can be observed when they present URL when voicing their view. In reply 2, user n4–211 inserts an URL after the claim “Eat it. Much better for you than seed oils and polysaturateds [URL],” without any elaboration of the content linked to by the URL, while in reply 4, user n4–2611 also posts a URL after stating the conclusion from the article linked to the URL, “British Nutrition Association ... recommended the low consumption of coconut oil ... and so far there’s no evidence of its health benefits [URL].” Significantly, user n4–1657 repeatedly posts the same URLs in replies 3 and 12 while rebutting the other two users’ posts. The practice of presenting URLs after their claim suggest that users employ URLs as evidence for their view. However, none of them details the content linked to, at best only the conclusions drawn. For example, the lack of elaboration in reply 2 attracts another user, n4–2611, to query if the user is “referring to coconut oil” in reply 4. The lack of elaboration points to the possibility that users only focus on the presence of URLs, rather than the information linked to, for their arguments, and may have taken URLs as “hard currency” for their stance (Wikgren, 2003).

Users using URLs as “hard currency” can also be observed when they respond to others with URLs. User n4–211 comes back in reply 5 with another URL to support their stance, “There is no problem with saturated [sic] fats,” in response to reply 4, introducing it with “see…..”, suggesting that URL is offered as evidence for their claim. This claim is aligned by user n4–1657 in reply 8 “It may well be the case that saturated fat is not that bad for you”, which is a conclusion made after the user shares another URL, “a very interesting article discussing the systematic reviews”. The exchange between these two users attracts user n4–1657’s strong objection in replies 9 and 10 and posting of a URL in reply 11 with a positive evaluative introductory frame, “Here is a good presentation on fat and health research”. Posting an URL at the end of one’s objection to others’ stance again suggests the use of URL as evidence for one’s stance.

Secondly, the frequent exchange of URLs among these three users means that URLs are elevated to be the main topic of the discussions. This focus of URL also renders the initiator’s response in reply 6 – that contains no URLs – irrelevant to the conversation, such that nobody picks up on this reply. More importantly, the focus on URLs hinders the negotiation process between users as they stick to their own URLs. As shown towards the end of this thread, the discussion evolves into a criticism and defense of URLs, as shown in reply 12, “Some of the quoted studies here are from 2000 and old research”, “Also these were funded by big pharma in view of supprting [sic] statin sales” and reply 13 rebuttal, “The newer the research, the bigger the commercial influence”. Although it is important to evaluate the credibility of sources, the mere focus on the credibility and URL itself to the point of side stepping the content may lead users into the situation of “He says, she says”, rather than really discussing what has been said. Coupled with users seldom elaborating on the content contained in the URLs posted by themselves, disagreeing users may not be able to understand what underlies each other’s point of views. Stalemate between the disagreeing users is evinced in reply 12 “Sorry not convinced.”, and reply 13 “If ….. does not convince you than nothing will”.

This thread illustrates a user-user interaction largely builds on a URL-URL interaction, such that the thread moves towards presentation, criticism and defense of URLs. Each user seems to take their own posted
URLs as the “hard currency” for their stance (Wikgren, 2003), such that the attention is on URLs, and there is little negotiation among users in their views, thus leading to stalemate towards the end. The obvious stalemate further points to the problem of reliance on URLs and link war. This URL exchange and the practice of posting URLs to argue against each other can only be observed if the whole thread is examined while taking into account users’ responses to each other. This observation would not be possible if only individual comments are considered. Furthermore, the implication of URL-posting and link war, in this case users’ singular focus on their own URLs that impedes on their negotiation with each other, is also revealed when the whole thread is examined. All these insights point to the importance of conducting micro-analyses of digital CA of threads in online discussions where a specific topic or language practice is salient. This example also shows that the CWB tool is useful for tabulating threads that contain most mentions of the search terms, which in turn helps researchers to focus on the threads that might be illustrative of a certain language practice.

Discussion and conclusion

In this article, I illustrated the use of CWB tool to explore a large corpus of online discussions and discussed how it could be used to facilitate a synergy of corpus analysis and interactional approach to provide insights into users’ interactions. CWB is stable for large corpora, thus suitable for storing and querying textual data of online discussions which usually contain thousands of comments. More importantly, CWB allows for annotations of variables and units of analysis regarding users’ interactions in the online discussions. This in turns allows for comparisons between subcorpora (keyword analysis) defined based on different variables (e.g., topics of discussions, role of internet users, types of postings) and concordance reading that could be expanded from lines to different units of analysis. Additionally, grouping search results at the level of threads facilitates extraction of threads containing many mentions of search terms that are of interest to researchers intending to conduct digital CA or micro-analysis.

I also argue for two methodological considerations in the corpus analysis of online discussions. Firstly, I propose that language practices in online discussions can also be explored based on different variables rather than assuming all postings are the same. For example, I have briefly presented language practices in initiating posts that trigger replies compared to those in the independent posts that do not receive replies. Interested readers can refer to Chua (2021) for a complete analysis. This kind of analysis will provide further insights into users’ interactions and language practices in online discussions.

Secondly, I showed that a synergy of corpus analysis and an interactional approach to online discussions can further our understanding of online interactions. As shown in the case studies, on top of the language practices revealed by the keyword analysis, micro-analysis further reveals the implications of language practices for users’ interactions. For example, the case studies reveal the concession strategy that facilitates negotiation among users and the URL exchange that does not seem to do so. It is of utmost importance to understand the impact of users’ language practices on online interactions given that online spaces nowadays seem to be susceptible to hostile interactions. Furthermore, conducting micro-analysis on threads containing keywords found in a corpus analysis also reveals how different keywords are used together to realize a specific language practice and how specific keywords are repeatedly used within one thread. As shown in the case studies, a concession strategy is realized by multiple keywords while users are found posting URLs repeatedly within a thread.

Undertaking these methodological considerations is the fact that users respond to each other in discussion threads. Only analyzing individual postings, without considering what it responds to or what comes after may decontextualize the postings, especially the replies, thus not providing a full insight into the language practices in online interactions. Nonetheless, this is not to say that concordance reading or corpus analysis are not suitable for investigating online discussions. Rather, the purpose of this article is to suggest a synergy of corpus analysis with an interactional approach, whether micro-analysis, digital CA or comparing keywords in different types of posts or posts contributed by different groups of users. Corpus analysis provides insight into language practices by analysing the big data available from online discussions while an interactional approach complements the findings of corpus analysis by delving deeper into users’ interactions. This is similar to the CLA approach proposed by O’Keeffe & Walsh (2012) for face-to-face interactions. Additionally, this paper also shows that corpus tools can be used to extract threads or provide a point of entry to data when conducting turn-by-turn analysis of threads in online discussions which typically contain large amounts of data.

Beside the MOOC online discussions investigated in this paper, online interactions, including Facebook comments, Twitter replies, and various online communities such as Quora and Reddit, could also be examined by drawing on both corpus linguistics and interactional approach. By combining these two approaches, public opinions and debates online could be better understood via the big data analysis and in-depth analysis of users’ interactions. This way, we could gain further insights into users’ interactions online, including disagreement, information exchange, propagation of misinformation. This is especially crucial in the current society where online space is also where people perform their social life. Other than online communications, this combination of approaches could also be applied to other interactional data, such as transcript of face-to-face conversations, while CWB tool can be used for big data requiring annotation of hierarchical structures.

The benefits, as well as the complexity, of integrating various research approaches, in this case corpus linguistics and interactional approach, also call for better integration of various research software for future development. Similar to other corpus tools, CWB does not have any functions for researchers to note down their analysis. In contrast, various qualitative data analysis software such as NVivo could have functions while having only rudimentary corpus analysis features. In the current analysis, I have employed CWB for corpus queries, then based on the results, used R to export threads in text files to be imported to NVivo so that I could note down my analysis on relevant part of the threads. Streamlining the query and analysis process would further improve researchers’ workflow and encourage exploration of a corpus.

This paper suggests various potential research questions that can be asked in relation to online discussions and how the CWB tool can facilitate the exploration and analysis such that researchers will not be constrained by the currently available corpus tools. However, researchers may not necessarily want to investigate the interactive aspects of online discussions that I have listed above and may have a different research focus, as shown by previous corpus analysis (e.g., Paterson, 2020; Sotillo and Wang-Gempp, 2016). It should also be noted that the configuration of CWB and encoding of the corpus requires some programming skills that might put some researchers off. Yet, once set up, the corpus, especially large corpus with more than 1 million words, can easily be accessed and analysed. Therefore, I recommend using CWB for the corpus analysis of online discussions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The author thanks Dr Caroline Tagg for advice on this manuscript and Leverhulme Trust Open World Learning Program for funding her PhD, part of which form the research presented in this paper.
References