

MONIKA BEDNAREK
THE UNIVERSITY OF SYDNEY

ANALYSING INTRATEXTUAL PATTERNS IN CORPUS-ASSISTED DISCOURSE STUDIES

CITATION

Bednarek, M. (2024). Analysing intratextual patterns in corpus-assisted discourse studies. *Journal of Corpora and Discourse Studies*, 7: 43-62

ABSTRACT

Many corpus-based discourse analysts are interested in linguistic patterns that hold across the texts in their corpus. In such corpus-based studies, relatively little attention is given to 'intratextual' patterns or discourse organisation/text structure. For example, a standard practice in corpus-assisted discourse studies is to generate lists of word frequencies, keywords, collocates, and n-grams on the basis of total word frequencies in the corpus, without analysing how these words are used to structure texts or conversations. In this commentary, I reflect on this issue, and introduce selected techniques for integrating analysis of text structure/discourse organisation in corpus-assisted discourse studies, including the analysis of dispersion/concordance plots in media data, the use of clusters/n-grams across sentence breaks in conversational data, and the repurposing of parallel concordances for analysing Twitter posts and responses to them.

KEYWORDS

Intertextual analysis; discourse organisation; text structure; corpus methods; dispersion

CONTACT

Monika Bednarek, Linguistics, The University of Sydney, Brennan MacCallum Building A18, Sydney, NSW 2006, Australia
monika.bednarek@sydney.edu.au

DOI

<https://doi.org/10.18573/jcads.113>

ORCID

0000-0001-7819-5424

ISSUE DOI

<https://doi.org/10.18573/jcads.v7>

LICENSE

© The authors. Available under the terms of the CC-BY 4.0 license

Analysing intratextual patterns in corpus-assisted discourse studies

Monika Bednarek
The University of Sydney

1. Introduction

In this commentary, I want to address the analysis of text structure/discourse organisation in corpus-assisted discourse studies and introduce selected techniques for integrating such analysis. Many corpus-assisted or corpus-based discourse analysts are interested in linguistic patterns that hold across the texts that are included in their corpus and often analyse such patterns at the level of the corpus rather than individual corpus texts (Egbert & Schnur 2018). In such studies, little attention is therefore given to the structure or discourse organisation ('intratextual' patterns) of individual corpus texts. For example, a standard practice in corpus-assisted discourse studies and corpus-based discourse analysis is to generate lists of word frequencies, keywords, collocates, and n-grams on the basis of the total frequency of these features in the corpus, without necessarily analysing how these features are used to structure texts. As a consequence, results may ignore variations within texts, thus neglecting or backgrounding text-internal structures. As Stubbs (2008, p. 5) notes with respect to keyword analysis, 'since the texts have been ripped apart into lists of individual words and/or n-grams, the patterns ignore text segmentation. They are a feature of global textual cohesion, but not textual structure.' Marchi & Taylor (2018, p. 4) point out that discourse analysts have long critiqued corpus linguists 'for mainly focusing on individual words and lacking insight into the various dimensions of discourse structure'. Disregarding or backgrounding text structure can be a tendency even for corpus linguistic studies that focus on dialogic text types such as question-answer forums (see e.g. most of the contributions to Baker & Egbert 2016).

Despite some notable exceptions, corpus-assisted discourse studies (CADS) of discourse organisation appear to have largely been limited to discussion of concepts such as the idiom principle/co-selection, cohesion, lexical priming, and semantic/evaluative prosody¹ (e.g. in various publications by Alan Partington and colleagues, such as Partington, 2004, 2014, 2017a; Partington *et al.*, 2013; Duguid & Partington, 2017; Morley & Partington, 2009, and others). When I say CADS, I am talking here specifically about the body of corpus-assisted discourse studies indebted to research projects carried out in Italy (Partington *et al.*, 2013: p. 10). As evident from this body of research, primings and pros-

1 For general debates around semantic prosody, see e.g. Whitsitt (2005), Hunston (2007), Bednarek (2008), Stewart (2010), and references cited therein.

odies can affect the immediate co-text of items as well as longer stretches of text (including across clause or sentence boundaries) and can also contribute to text cohesion.

This focus may be different in other studies that combine corpus linguistics with discourse analysis and that we could perhaps label as CADS in a very broad conceptualisation of the field, for instance, studies in corpus pragmatics, corpus-based critical discourse analysis (CDA), or corpus-based register analysis.² On the overlaps between various approaches to corpora and discourse that use different labels, see e.g. Marchi & Taylor (2018, p. 5). My aim here is not to boundary-police CADS as a field, but rather to enable a sharper focus to the present commentary.

Honing in on studies self-identifying as CADS, a search for “corpus-assisted discourse studies” retrieves 2,060 results on Google Scholar (as of 2023-07-17). Adding expressions such as “discourse organisation”, “text structure”, “discourse structure”, “conversational structure”, or “conversational organisation” retrieves fewer results by far (ranging from two to 139). While this is a relatively crude method, it does indicate a trend within CADS research to background text structure or discourse organisation. Given the scope of this short piece, I refrain from giving an overview of CADS or other corpus linguistic studies that *have* integrated analysis of text structure/discourse organisation, but Flowerdew (2005) reviews genre-based corpus studies, Barlow (2016) reviews corpus studies that look at specific text segments, and Schnur & Csomay (2020) review computational and manual approaches to text segmentation. There was also a project focusing specifically on structure within news discourse (e.g. O’Donnell *et al.*, 2012) that may be of particular interest to CADS scholars.

In this commentary, I introduce selected techniques for integrating analysis of text structure/discourse organisation in CADS. These techniques include the analysis of dispersion/concordance plots, clusters/n-grams across sentence breaks, and the repurposing of parallel concordances. Many of the points I make in this commentary also apply to corpus-based CDA (see Narthey & Mwinlaaru, 2019) and should therefore be of relevance beyond CADS.

2 See e.g. some of the contributions to Partington *et al.* (2004) on discourse organisation, discourse signposting, and pragmatics, Flowerdew & Mahlberg (2006), or Schnur & Csomay (2020) to name but a few. An overview of some pertinent corpus linguistic research on interactive spoken discourse is provided in Partington *et al.* (2013, p. 210–213). A potentially interesting novel technique for analysis of cohesion – correlation analysis – is discussed in Brookes & McEnery (2020), and there are other approaches used within corpus linguistics as a whole or in fields adjacent to corpus linguistics such as natural language processing/computational linguistics, including but not limited to work on coherence relations (e.g. Das & Taboada, 2018).

2. A topology

As a background to this discussion, I will use the topology in Figure 1, a modification of an earlier version introduced in Bednarek & Caple (2017a, b). The term *topology* refers to scalar rather than categorical distinctions which are represented in taxonomies (Martin & Matthiessen, 1991). The topology was developed to allow researchers to both explicitly situate their study (for transparency) and to increase researcher reflexivity, i.e. ‘a greater consideration of how the researcher impacts on the researched’ (Baker, 2012, p. 255). In other words, it aims to encourage ‘self-reflexivity’ (Marchi & Taylor, 2018, p. 12).

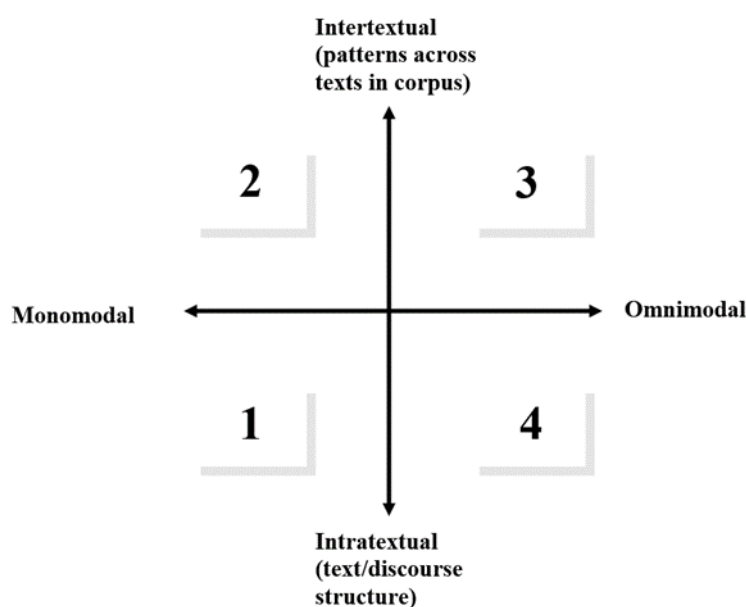


Figure 1: A topology for situating research

In essence, the topology focuses on (scalar) distinctions between different foci of research. One of these distinctions concerns the researcher’s approach to mode. The other, which is most relevant to this commentary, focuses on the researcher’s approach to texts. The corresponding distinctions are:

- a) Horizontal axis: whether the primary focus of the study is on one semiotic mode (is monomodal) or more than one (is multimodal or even omnimodal, analysing all possible modes);³

3 Strictly speaking, texts that are ‘multimodal’ combine two or more modalities (e.g. visual, aural), whereas ‘multi-semiotic’ texts combine two or more semiotic (meaning-making) systems such as image or language (O’Halloran, 2008). However, the term multimodal has typically been employed to mean both. Much CADS research is monomodal, but this is a matter I do not address in this commentary (see e.g. Caple, 2018 for detailed discussion).

- b) Vertical axis: whether the primary focus of the study is on patterns across texts (intertextual) or patterns within texts (intratextual; text structure/discourse organisation).

These axes intersect, producing zones of analysis where we can situate studies that are e.g. simultaneously monomodal and intratextual (zone 1).

It is important to note that this use of the terms *intratextual* and *intertextual* differs from other uses in corpus linguistics, where *intratextual* analysis may refer to the analysis of a single text, and *intertextual* analysis may refer to analysis of allusions and similar phenomena (e.g. Adolphs, 2006, p. 66–69). Intertextual patterning is also different from transdiscursive intertextuality in CADS (Duguid & Partington, 2017) and from how intertextuality is typically defined in CDA (broadly, as links between texts).

To clarify, studies interested in intertextual patterns focus on patterns or trends across the texts in a corpus, while studies interested in intratextual patterns focus on the development of meaning within individual texts (e.g. generic stages, discourse moves, turn-taking patterns, speech act sequences, etc.) - ‘the unfolding of meaning in texts’ (Bednarek, 2009, p. 20). Some prototypical examples of an intratextual study would be the analysis of turn-taking sequences in a conversation among friends, the analysis of cohesion in a political speech, or the analysis of moves in a research article. Many corpus linguistic studies, including CADS, are positioned in zone two of this topology: they tend to be both monomodal and intertextual. Another trend, discussed in Egbert & Schnur (2018), is that findings from such studies are typically reported for the whole corpus, thus ignoring distribution or dispersion across individual texts.

The point of this topology is not necessarily to help researchers design a study that covers everything, but rather to make them aware of what aspects of the data are not captured in the analysis. In other words, the topology can be useful for situating or positioning specific corpus studies in terms of their approaches to the analysis of semiotic features within and across the texts in their corpora. I will draw on the topology’s distinction between intertextual and intratextual analysis in the next section, which discusses existing corpus techniques/tools for analysis of text/discourse structure.

3. Analysing intratextual patterns using existing corpus techniques/tools

In general, it appears that there is a relative lack of corpus tools that facilitate corpus-based analysis of intratextual patterns. As explained in section 2, by this I mean examining the patterns that hold *within* the texts included in a corpus, such as turn-taking structure (if conversational data are examined) or move structure (in different text types/genres). Thus, Schnur & Csomay (2020: 22) point out that ‘[r]esearchers interested in examining issues of text structure have often had difficulty reconciling this goal with cor-

pus-based techniques' and that existing corpus tools 'are generally ill-equipped to examine [...] intra-textual features that span larger sections of text.' It appears that the situation is similar regarding information visualisation tools (Martin-Rodilla & Sánchez, 2020). As Egbert & Schnur (2018: 161) further note, 'existing software makes it difficult or impossible to visualise and describe findings from a corpus at the level of the text.'

Despite this general trend, there are some existing corpus techniques/tools that are currently available for analysis of intratextual patterns, even if their features may be rather limited. Given the scope of this commentary, I will not offer a systematic or comprehensive summary of all potentially-relevant CL tools/techniques, but rather focus on a selection that I have applied in my own research and that I consider to be somewhat underutilised in CADS. I will therefore *not* discuss cases where particular terms/phrases uttered by speakers are the starting point(s) for analysis, and where subsequent analysis then shows how they function in interactions,⁴ a workflow also used outside CADS in contexts such as business communication (e.g. Koester & Handford, 2018 on hypothetical reported speech within sequential patterns). It is clear that this workflow (or technique) enables analysis of discourse-organising features and their use in texts. However, this technique of "shunting" between concordancing and close reading' or text analysis is very well-known in CADS (Partington & Marchi, 2015, p. 231). Moreover, I will *not* discuss the analysis of corpora that are annotated for speakers, speaker turns, text structure, discourse units (e.g. Egbert *et al.*, 2021), and so on. Segmentation of language into self-contained textual units is a challenge in itself and different text segmentation methods exist (e.g. Egbert & Schnur, 2018, p. 163-164). If included in the corpus annotation (e.g. tagging answers/questions in an answer-question forum corpus; see Baker & Egbert 2016), one can at least theoretically undertake CADS analyses that take into account speaker interaction or text structure. The latter can also be achieved when different sub-corpora are compiled which correspond to different text-structural units (such as a corpus of advice-seekers' questions and a separate corpus of corresponding answers; see Carr 2020). Since automated text segmentation tools in corpus linguistics are summarised in Schnur & Csomay (2020), section 4 describes a range of other corpus tools/techniques that can be used to examine intratextual patterns, drawing on examples from my own research.

4. Intratextual analysis: Analysis of text structure

One of the tools available to CADS researchers who wish to undertake intratextual analysis is dispersion plot analysis. This allows users to inspect how linguistic features are distributed within the individual files that make up a corpus. Dispersion plots (also called concordance plots) are built into offline tools such as AntConc (Anthony, 2019) and

4 See for example Partington (2008, p. 110) on the use of *if I may/can* in press briefings. These briefings are also explored in other publications by Alan Partington, for example in relation to 'forced priming' (Partington *et al.*, 2013, p. 209).

WordSmith Tools (Scott, 2019) and are also included in many online tools. They show where a search term appears and clusters in individual files. Dispersion plot analysis has in fact been used in CADS (e.g. Partington *et al.*, 2013, p. 253, p. 257), but I would argue that it is still under-utilised. An example study using dispersion plot analysis is Bednarek (2015), which applies the tool to explore identity construal via pronouns in a scientific radio programme, tracing their textual patterning to show changes from audience address to individual subjectivity and community alignment. Rather than using a single dispersion plot, different plots are aligned to show changes in a single text (see Figure 2), but this could be extended to other texts in a corpus. Another example is provided in Bednarek (2018, p. 165–169), which illustrates the use of dispersion/concordance plots for examining the interaction between key words and plot development in fictional texts.

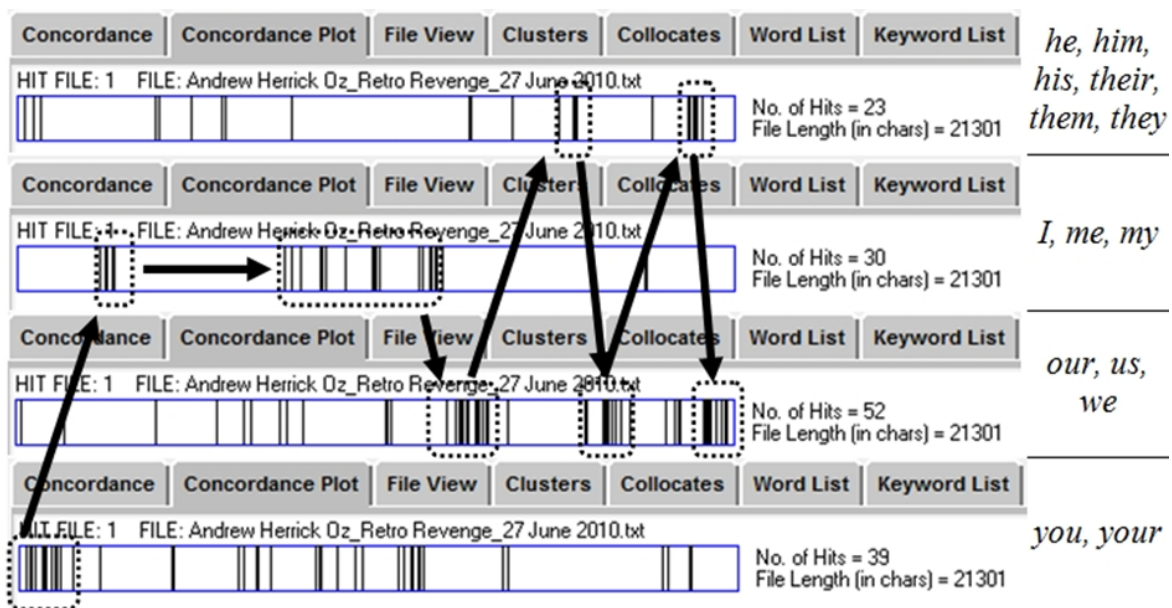


Figure 2: Identity construal via pronouns

Other uses are also possible, and dispersion plot analysis could be used to investigate (i) whether specific linguistic features occur in certain stretches of texts and (ii) whether this trend holds across corpus files. Such a study would *simultaneously* be positioned in zone one and two of the topology in Figure 1. With such analysis it is important to ensure that corpus files correspond to meaningful textual units. For example, is each file recognisably self-contained and functional (Egbert & Schnur, 2018: 162)? Where corpora contain random text fragments or where one file consists of multiple textual units, dispersion analysis of these files is not likely to provide us with much meaningful information about intratextual patterns.⁵

5 If the file contains a whole corpus (all corpus texts are included in one file), the plot visualises the distribution across the corpus as a whole. Dispersion can also be visualised across corpus sub-categories (for discussion, see e.g. Anthony, 2018, pp. 213–215).

Another tool for integrating analysis of text structure/discourse organisation is the analysis of clusters/n-grams *across* sentence breaks with the aim of describing the patterns that hold across the sentences in a text (e.g. turn-taking patterns). This technique hence identifies intratextual patterns across sentence boundaries rather than longer text stretches. As an example, Bednarek (2010, 2014) used cross-sentence n-gram analysis to identify ‘interactive clusters’ (Bednarek, 2010, p. 73) which are distributed across speaker and addressee (e.g. *how are you good; why not because*) and allow the observation of recurring patterns of interaction (e.g. *Why not? -Because... = demand for explanation + account*). Concordance analysis is necessary to confirm whether a particular cluster/n-gram is distributed across different speakers and is indeed used to structure the conversation. One could then use range analysis (Nation & Waring, 1997) to identify in how many corpus texts a particular interactive cluster is used, thus combining intratextual with intertextual analysis – with range analysis focussing on distribution across corpus files. It is worth noting that corpus tools differ in their default settings for clusters/n-grams, i.e. whether n-grams are by default retrieved across sentence boundaries or not. To be able to decide when it is appropriate to respect sentence boundaries and when it is appropriate to focus on cross-sentence clusters/n-grams, it is crucial that CADS scholars are aware of default settings and possibilities for adjustment (for instance, via changes made in the relevant tool settings for n-grams retrieval).

It is also possible to repurpose corpus tools that were not originally developed for intratextual analysis. An example is Fuoli & Bednarek’s (2022) use of the parallel concordance tool (originally developed for multilingual corpora) in SketchEngine (Kilgarriff *et al.*, 2004) for analysis of dialogic patterns in Twitter (now X) data, namely in a corpus of 1300 complaint-response interactions from 13 airlines between March and July 2020. A parallel concordance shows the original text in one language aligned with its translation into one or more other languages. However, the tool can be repurposed such that responses to particular Twitter posts can be shown instead, i.e. the customer’s tweets are presented as ‘original’ text and agents’ responses are treated as ‘translated’ text. This requires prior alignment of relevant utterances (customer tweets and agents’ responses). An example from Fuoli & Bednarek’s corpus (but not used in their article) is shown in Table 1, which includes all instances of the lemma BAD (*bad, worse, worst*) in the complaining tweets together with the relevant responding tweets.

Concordance	Original text	Agents' responses
1	@ [airline] OH MY GOD! ANA! i just cannot believe customer service never reply my email nor pick up my phone call for more than 1 week! It is unbelievable that luxury airline refund service is worse than cheaper airline like [airline]!!	@ [name] Good day, @ [name]. As we've mentioned in our previous responses, it may take some time for you to be able to connect with our Reservations and Customers Service Center as they are also receiving a high volume of inquiries due to the recent events that are happening.
2	Trying to cancel our flights to Japan with @ [airline]. Have been on hold for an hour. No sign of progress. The worst part of it is that we dialed from my partner's phone and he has to go to work so we have to hang up	@ [name] We sincerely apologize for the inconvenience, @ [name]. However, please be advised that our Reservation and Customer Service Center is experiencing high call volume due to the influx of inquiries they receive, and it may take a while to successfully connect. You may also try to reach them at [phone number] (charged). Thank you, and we hope this helps.
3	Awful @ [airline] happily took my money on March 17 to rebook my elderly mother on a May 5th flight home from Oz (instead of March 28 during the worst of international/airport covid spread) . Now they've cancelled her flight and said they will do nothing to help get her home. [url]	@ [name] Hello [name], could you please DM her booking details? We'll take a look into it. /[name]
4	@JustinTrudeau announced today to all canadians outside of Canada to come back to Canada as soon as we can. we're here in Bogota-Colombia trying to contact @ [airline] but it's impossible. more that 7 hours in the phone and not even one answer from anyone. worse costumers service.	@ [name] Hello [name], we are sorry to learn about this and invite you to send us a DM with your booking details so we may better assist. /[name]
5	@ [airline] The only way to use my	@ [name] Hi, [name]. I'm sorry you

Concordance	Original text	Agents' responses
	voucher is by calling you, but all of your #CustomerService call centres are unavailable. You have Morphe d from one of my favourite #airlines to the worst . Ridiculous! [url]	can't use your voucher to rebook online. It's something we're looking at, but we can't say when it will be possible. However, we can help you, but you will need to follow us and send us your details via DM. [name] [url]
6	@ [airline] Worst airline ever. No info from [airline] despite travel changes, the customer have to do their job. I'm not used to go through social media, but you fairly earned that msg. You're a shame. If you have to travel with [airline], don't expect any professional behavior. Be proactive.	@ [name] We've already issued special ticketing arrangement ([url]) and notified all of the affected passengers. If tickets are issued with travel agent, we might not be able to reach passengers directly. You can DM us for further help if tickets are issued with us. -[initial] [url]
7	@ [airline] sent the refund rqst proof from March when I initially requested for the full refund in CASH before covid19 got worse . And now all of sudden I got the email it says [airline] generated travel voucher this is again not acceptable please check dm and release my payment	@ [name] We have already responded to your DM.
8	This is too bad and frustrating! @ [airline] how can I pay for same flight twice, and I did not a ticket or refund. Tried all possible best to get in touch all to no success. Real Scam!	@ [name] We have replied to your DM.
9	@ [airline] Flight cancelled March,dealings on messenger,emails,full refund requested and still awaiting a refund! Worst customer services ever..	@ [name] We would gladly assist you with this. Can you please send us your booking code in a DM?
10	@ [airline] You offer me a 8\$ Voucher for a ticket you canceled on me in march i paid 226.52€ for? I hope this is a bad joke. Would you care to explain?	@ [name] We've verified the refund, and it's still in process. In this case you must wait to receive the new travel voucher. Furthermore, we suggest you

Concordance	Original text	Agents' responses
	[url]	to delete the images, due to it contains private information. Don't hesitate to write us back in case you need our help. Regards.
11	@ [airline] Now the app and website are functioning again. But really bad I cannot pay with an [airline] voucher I received today. When will there be an option in the payment screen to use the voucher code? Don't want to call again. Last week on the fone with [airline] for 30 minutes. Cost me \$\$\$\$\$	@ [name] You can do a claim to get assistance by email in the next link: [url], let us know if you have more questions. Regards.
12	@ [airline] I am horribly disappointed in [airline] and their handling of my refund. This has been the worst customer experience of my life and after 100 hours on the phone in the past month and all of this effort I still do not have my money back or an option to rebook my flight	@ [name] Unfortunately, we are unable to provide an exact timeframe as to when you will receive your refund. Please rest assured that we have increased capacities and optimized procedures so that you may receive your refund as quickly as possible. / [name]
13	@ [airline] How do you get in contact with you guys 1 1/2 hours waiting. Seriously bad service and what other options are there. You changed my flights and my only options were to accept or take a credit. Also your music is awful. I would love a response.	@ [name] Hi [name], appreciate your patience with our long wait times due to the influx of Coronavirus related enquiries we are receiving. Please DM us your query so we can best assist. [initial]
14	@ [airline] I bought a ticket to Bali before the Covid19 outbreak but then my flight was cancelled when things got bad . Now I have attempted contacting [airline] about getting a refund or at least travel credit but I have heard nothing, [airline] please help me.	@ [name] Hi [name], thank you for getting in touch. Please feel free to message us on here with your booking reference number so we can check this further for you. [name] [url]
15	@ [name] @ [airline] That is not fair!!	@ [name] @ [name] Hello [name],

Concordance	Original text	Agents' responses
	At least you were able to get in touch with them. For the past 3 days no one is attending my call very bad customer service. If they cannot handle this unprecedented situation, no need to flex their arms on promoting flights this period	Our call centres are experiencing a high volume of calls at this time and we apologize if we cannot immediately answer your call. You may please share your query with us on Direct message, We will certainly assist you in the best possible way. Thank you.
16	@ [airline] I have requested several refunds for my flights in March and April and after a very long waiting time you have sent me vouchers and there is no option to get in touch with you and ask for a cash refund. Very bad customer service @ [airline]	@ [name] Hi, can you please DM us your reference number, passenger names, route, dates and email used for the booking. I can then add you to the refund queue.
17	@ [airline] You are holding my group deposit since 5 months now and not giving me refund. There is no chance my guests can travel to Australia this year. But you are treating us like a piece of shit and not processing our refund. Your local Ahmedabad office is telling it as HO POLICY. Worst .	@ [name] Hi [name], refunds are being processed in the order of the earliest scheduled dates of departure. A notification will be sent once your request has been processed. We ask for your patience as this may take longer than usual, due to the backlog of service requests arising from Covid-19. Thank you.
18	@ [airline] won't refund my money!!!! Keep getting the run around!!!! Worst airline ever!!! They won't refund me over 5k in flights!!!	@ [name] Hi there, as long as your ticket meets the eligibility criteria of our enhanced travel waiver policy([url]), you may request a refund. If you have made your booking directly with [airline], you may submit your request via [url]. If you have made your booking with a travel agent, please contact them for assistance. Thank you.

Table 1: Parallel concordance for the lemma BAD

Fuoli & Bednarek (2022) illustrate how such concordance displays can be used to explore patterns of dialogic interaction between Twitter users, in their case with an emphasis on *emotional labour* (Hochschild, 1983). However, it is clear that this technique has applications for CADS that go beyond this topic in studies where the analysis concerns utterance pairs such as complaint-response, question-answer, or other sequences of two speech acts. While SketchEngine requires paid access, the Australian Text Analytics Platform (ATAP) team at the University of Sydney is developing a Concordancer in the form of a Jupyter notebook which enables similar analysis. Figure 3 shows a screenshot from the beta version of the notebook (Bednarek *et al.*, 2023) with three Tweets containing the word *ridiculous* aligned with their relevant responses (in the ‘text’ and ‘response’ columns on the right). The notebook still requires further development but is a proof-of-concept example that users are able to test with small datasets.

The screenshot shows the ATAP Concordancer interface. At the top, there is a search bar with the term 'ridiculous'. Below it are checkboxes for 'Enable regular expressions' (unchecked), 'Ignore case' (checked), and 'Match whole words' (checked). There are also dropdown menus for 'Page' (1), 'Window size (characters)' (50), and 'Sort by' (text_id). A 'Show More ...' dropdown is open, showing 'text' and 'response' options. Below the search controls is a 'Filename (without.xlsx extension)' input field and an 'Export to Excel' button. The main part of the interface is a table with the following columns: left_context, match, right_context, text_id, text, and response. The table displays three rows of results, each corresponding to a tweet containing the word 'ridiculous' and its response.

left_context	match	right_context	text_id	text	response
s beyond nervous 0 last time I flew, it was sort of	ridiculous	. Do the right thing please, I have an upcoming fl	49	@ with Covid-19 cases rising again. I want to ask you to do the right thing and not sit the middle seats on your flights. I was beyond nervous last time I flew, it was sort of ridiculous. Do the right thing please, I have an upcoming flight.	@ Hi, Ray. We're looking forward to seeing you again. Check out all that we're doing under Your Travel Experience. https://
1 What a @ put all your t	ridiculous	move made for profit. @ put	60	What a ridiculous move made for profit. @ put all your top execs on these crammed flights as personnel. https://	@ We've implemented many layers of protection including enhanced cleaning and requiring face coverings. https://
2 @ @ This is	ridiculous	already... My flight has been switched from AirFr	105	@ @ @ This is ridiculous already... My flight has been switched from AirFrance >> KLM to leave days later, missing best friends wedding! no one is answering/ being helpful!	@ Hello @ we regret we were unable to reply to you faster. Do let us know if our assistance is still required, please provide us with your booking reference via DM or renew your request. We remain available.

Figure 3: A screenshot from ATAP's Concordancer Jupyter notebook, aligning complaints (*text*) and replies (*response*); without user names and URLs

There are other corpus-based tools/techniques that can assist in intratextual analysis, but such tools are arguably still in their infancy. One of the oldest tools of this type is AntMover (Anthony, 2003) for move-like analysis, but it has seen little development since its initial release. There are also several other move analysis tools, such as AWSuM (Mizumoto *et al.*, 2017), but these are built on pre-annotated corpora. Another interesting tool is WordSkew (Barlow, 2016), which identifies the frequency of a search term de-

pending on where it appears in individual texts based on html/sgml markers and some automatic recognition (e.g. of sentence/paragraph boundaries). It was specifically designed to address the linking of corpus data with discourse structure. The Concord tool in WordSmith also allows some analysis of the position of search terms in specific text segments (sentences, paragraphs, headings, sections). Another example is provided in the CLiC tool (Mahlberg *et al.*, 2016), which allows users to explore the distribution of features in particular segments of literary texts (such as within quoted speech or in suspensions—stretches of narrator text interrupting character speech). Another recently-developed Jupyter notebook by the University of Sydney's ATAP team, the Quotation Tool (Jufry & Sun 2022), enables the construction and analysis of English-language newspaper corpora distinguishing between quoted vs non-quoted content. Tools in disciplines adjacent to corpus linguistics may also exist, e.g. in computational linguistics, computer science, or in interdisciplinary fields such as Digital Humanities and Computational Social Science. Discursis (e.g. Angus *et al.* 2012) is but one such example, which offers automated analysis and visualisation of participant interactions around topics over time. It has recently been reworked as an open-source tool through the ATAP team, with this version of the tool available on GitHub (<https://github.com/Australian-Text-Analytics-Platform/discursis>).

5. Concluding remarks

Corpus-assisted discourse studies often tend to analyse language patterns across texts rather than within texts, especially when media corpora are analysed. There is, however, a great potential for CADS to explore intratextual patterns and for corpus software developers to make innovations in order to assist researchers in this endeavour. Such innovations may then in turn lead to the discovery of 'non-obvious meaning' (Partington, 2008) at the level of text structure/discourse organisation, i.e. 'at the more macro-level of (non)obviousness on the plane of discourse' (Partington, 2017b, p. 339). To push CADS forward, we should continue to examine how we can best harness corpus linguistic techniques for exploring the flow of discourse (and discourses) within texts. This includes both developing new tools/techniques as well as adapting, adjusting or repurposing existing tools/techniques. Given the differing skillsets and institutional resources that CADS scholars may have at their disposal, there is no one-size-fits-all solution that I would want to prescribe here. However, I hope that this commentary has at least given researchers some food for thought and will lead to future developments in the corpus-assisted study of intratextual patterns.

Acknowledgements

This article was conceived after conversations with Laurence Anthony and Jesse Egbert, and I am grateful for their input into various draft versions. Figure 2 has been reproduced (with permission) from the chapter ‘What we contrarians already know’: Individual and communal aspects of attitudinal identity’ (Bednarek 2015) published in Maggie Charles, Nicholas Groom & Suganthi John (Eds.) (2015), *Corpora, grammar and discourse*. In honour of Susan Hunston (pp. 257-281). Amsterdam/Philadelphia: John Benjamins. The Concordancer and Quotation Tool Jupyter notebooks and relevant python scripts were developed by the Sydney Informatics Hub in collaboration with the Sydney Corpus Lab for the Australian Text Analytics Platform (<https://www.atap.edu.au>; <https://doi.org/10.47486/PL074>).

Funding

This project received investment from the Australian Research Data Commons (ARDC), which is funded by the National Collaborative Research Infrastructure Strategy (NCRIS).

Competing interests

The author has no competing interests to declare.

References

- Adolphs, S. (2006). *Introducing electronic text analysis. A practical guide for language and literary studies*. London/New York: Routledge.
<https://doi.org/10.4324/9780203087701>
- Angus, D., Smith, A. E., & Wiles, J. (2012). Human communication as coupled time series: Quantifying multi-participant recurrence. *IEEE transactions on audio, speech, and language processing*, 20(6), 1795-1807.
<https://doi.org/10.1109/TASL.2012.2189566>
- Anthony, L. (2003). *AntMover (Version 1.1.0)*. Tokyo, Japan: Waseda University.
<https://www.laurenceanthony.net/software>
- Anthony, L. (2018). Visualisation in corpus-based discourse studies. In Taylor, C. & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 197-224). London/New York: Routledge. <https://doi.org/10.4324/9781315179346-10>
- Anthony, L. (2019). *AntConc (Version 3.5.8)*. Tokyo, Japan: Waseda University.
<https://www.laurenceanthony.net/software>

- Baker, P. (2012). Acceptable bias? Using corpus linguistics methods with Critical Discourse Analysis. *Critical Discourse Studies*, 9 (3), 247-256.
<https://doi.org/10.1080/17405904.2012.688297>
- Baker, P. & J. Egbert (Eds.) (2016). *Triangulating methodological approaches in corpus linguistic research*. London/New York: Routledge.
<https://doi.org/10.4324/9781315724812>
- Barlow, M. (2016). WordSkew: Linking corpus data and discourse structure. *International Journal of Corpus Linguistics*, 21(1), 104-114.
<https://doi.org/10.1075/ijcl.21.1.05bar>
- Bednarek, M. (2008). Semantic preference and semantic prosody re-examined. *Corpus Linguistics and Linguistic Theory*, 4(2), 119-139.
<https://doi.org/10.1515/CLLT.2008.006>
- Bednarek, M. (2009). Corpora and discourse: a three-pronged approach to analysing linguistic data. In M. Haugh (Ed.), *Selected proceedings of the 2008 HCSNet workshop on designing the Australian National Corpus* (pp. 19-24). Somerville, Massachusetts: Cascadilla Proceedings Project.
<http://www.lingref.com/cpp/ausnc/2008/index.html>
- Bednarek, M. (2010). *The language of fictional television*. London/New York: Continuum.
- Bednarek, M. (2014). Involvement in Australian talkback radio: A corpus linguistic investigation. *Australian Journal of Linguistics*, 34(1), 4-23.
<https://doi.org/10.1080/07268602.2014.875453>
- Bednarek, M. (2015). 'What we contrarians already know': Individual and communal aspects of attitudinal identity. In M. Charles, N. Groom & S. John (Eds.), *Corpora, grammar and discourse. In honour of Susan Hunston* (pp. 257-281). Amsterdam/Philadelphia: John Benjamins. <https://doi.org/10.1075/scl.73.12bed>
- Bednarek, M. (2018). *Language and television series*. Cambridge: CUP.
<https://doi.org/10.1017/9781108559553>
- Bednarek, M., Mather, M., Maras, K., & Croser, H. (2023). ATAP Concordancer (v0.5.4) [Computer software].
https://github.com/Australian-Text-Analytics-Platform/atap_widgets. DOI:
<https://doi.org/10.5281/zenodo.10146967>

- Bednarek, M. & H. Caple (2017a). *The discourse of news values*. Oxford: OUP.
<https://doi.org/10.1093/acprof:oso/9780190653934.001.0001>
- Bednarek, M. & H. Caple. (2017b). Introducing a new topology for (Multimodal) Discourse Analysis. In P. Chappell & J. S. Knox (Eds.), *Transforming contexts. Papers from the 44th International Systemic Functional Congress*. Wollongong: 44th ISFC Organising Committee, <https://www.isfc2017conference.com/copy-of-sfl-lct>
- Brookes, G., & McEnery, T. (2020). Correlation, collocation and cohesion: A corpus-based critical analysis of violent jihadist discourse. *Discourse & Society*, 31(4), 351-373. <https://doi.org/10.1177/0957926520903528>
- Caple, H. (2018). Analysing the multimodal text. In C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 85-109). London/New York: Routledge. <https://doi.org/10.4324/9781315179346-5>
- Carr, G. (2020). 'Is this normal?': Examining sex education in a corpus of magazine advice columns. *Australian Review of Applied Linguistics*, 43(2), 145-168.
<https://doi.org/10.1075/aral.00030.car>
- Das, D. & M. Taboada (2018). RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52(1), 149-184.
<https://doi.org/10.1007/s10579-017-9383-x>
- Duguid, A. & A. S. Partington (2017). Forced lexical primings in transdiscursive political messaging. How they are produced and how they are received. In M. Pace-Sigge & K. J. Patterson (Eds.), *Lexical priming. Applications and advances* (pp. 67-92). Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.79.03dug>
- Egbert, J. & E. Schnur (2018). The role of the text in corpus and discourse analysis. Missing the trees for the forest. In C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 159-173). London/New York: Routledge.
<https://doi.org/10.4324/9781315179346-8>
- Egbert, J., Wizner, S., Keller, D., Biber, D., McEnery, T. & Baker, P. (2021). Identifying and describing functional discourse units in the BNC Spoken 2014. *Text & Talk*, 41(5-6), 715-737. <https://doi.org/10.1515/text-2020-0053>
- Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. *English for Specific Purposes*, 24(3), 321-332.
<https://doi.org/10.1016/j.esp.2004.09.002>

- Flowerdew, J. & M. Mahlberg (Eds.) (2006). Special issue of *International Journal of Corpus Linguistics*, 11(3) on 'Lexical cohesion and corpus linguistics'.
<https://doi.org/10.1075/ijcl.11.3>
- Fuoli, M. & M. Bednarek (2022). Emotional labor in webcare and beyond: a linguistic framework and case study. *Journal of Pragmatics*, 191, 256-270.
<https://doi.org/10.1016/j.pragma.2022.01.016>
- Jufri, S. & C. Sun (2022). Quotation Tool. v1.0. Australian Text Analytics Platform. Software. <https://github.com/Australian-Text-Analytics-Platform/quotation-tool>
- Koester, Almut & Handford, Michael (2018). 'It's not good saying 'Well it it might do that or it might not': hypothetical reported speech in business meetings. *Journal of Pragmatics*, 130, 67-80. <https://doi.org/10.1016/j.pragma.2018.03.005>
- Hochschild, A. (1983). *The managed heart: The commercialization of human feeling*. Berkeley, CA: University of California Press.
- Hunston, S. (2007). Semantic prosody revisited. *International Journal of Corpus Linguistics*, 12, 249-268. <https://doi.org/10.1075/ijcl.12.2.09hun>
- Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The sketch engine, *Proceedings of the 11th EURALEX International Congress*, 105-116.
- Mahlberg, M., Stockwell, P., de Joode, J., Smith, C., & O'Donnell, M. B. (2016). CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics. *Corpora*, 11(3), 433-463. <https://doi.org/10.3366/cor.2016.0102>
- Marchi, A. & Taylor, C. (2018). Introduction: Partiality and reflexivity. In Taylor, C. & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 1-15). London/New York: Routledge. <https://doi.org/10.4324/9781315179346-1>
- Martin-Rodilla, P., & Sánchez, M. (2020). Software support for discourse-based textual information analysis: A systematic literature review and software guidelines in practice. *Information*, 11(5), 256. <http://dx.doi.org/10.3390/info11050256>
- Martin, Jim R. & Christian M. I. M. Matthiessen (1991). Systemic typology and topology. In F. Christie (Ed.), *Literacy in social processes* (pp. 345-383). Darwin: Centre for Studies in Language in Education.
- Morley J. & Partington A. (2009). A few frequently asked questions about semantic-or evaluative-prosody. *International Journal of Corpus Linguistics*, 14, 139-158.
<https://doi.org/10.1075/ijcl.14.2.01mor>

- Nartey, Mark & Mwinlaaru, Isaac N. (2019). Towards a decade of synergizing corpus linguistics and critical discourse analysis: A meta-analysis. *Corpora*, 14(2), 203-235. <https://doi.org/10.3366/cor.2019.0169>
- Nation, I.S.P. & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6-19). Cambridge: Cambridge University Press.
- O'Donnell, M., Scott, M., Mahlberg, M. & Hoey, M. (2012). Exploring text-initial words, clusters and concgrams in a newspaper corpus. *Corpus Linguistics and Linguistic Theory*, 8(1), 73-101. <https://doi.org/10.1515/cllt-2012-0004>
- O'Halloran, K. L. (2008). Multimodality around the world: Past, present, and future directions for research. 35th International Systemic Functional Congress (ISFC), Sydney, 21-25 July 2008.
- Partington, A. (2004). 'Utterly content in each other's company'. Semantic prosody and semantic preference. *International Journal of Corpus Linguistics*, 9, 131-15. <https://doi.org/10.1075/ijcl.9.1.07par>
- Partington A. (2008). The armchair and the machine: Corpus-Assisted Discourse Studies. In C. Taylor Torsello, K. Ackerley & E. Castello (Eds.), *Corpora for university language teachers* (95-118). Bern: Peter Lang.
- Partington, A. (2014). Evaluative prosody. In K. Aijmer & C. Rühlemann (Eds.), *Corpus pragmatics: A handbook* (pp. 279-303). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139057493.015>
- Partington, A. (2017a). Evaluative clash, evaluative cohesion and how we actually read evaluation in texts. *Journal of Pragmatics*, 117, 190-203. <https://doi.org/10.1016/j.pragma.2017.06.008>
- Partington, A. (2017b). Varieties of non-obvious meaning in CL and CADs: from 'hindsight post-dictability' to sweet serendipity. *Corpora*, 12(3), 339-367. <https://doi.org/10.3366/cor.2017.0124>
- Partington, A., Duguid, A. & Taylor, C. (2013). Patterns and meaning in discourse. *Theory and practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam/Philadelphia: John Benjamins. <https://doi.org/10.1075/sci.55>
- Partington, A. & Marchi, A. (2015). Using corpora in discourse analysis. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 216-234).

- Cambridge: Cambridge University Press.
<https://doi.org/10.1017/CBO9781139764377.013>
- Partington, A., J. Morley & Haarman, L. (Eds.) (2004). *Corpora and discourse*. Bern: Peter Lang.
- Marchi, A. & Taylor, C. (2018). Introduction: Partiality and reflexivity. In C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review*. (pp. 1-15). London/New York: Routledge. <https://doi.org/10.4324/9781315179346-1>
- Mizumoto, A., Hamatani, S., & Imao, Y. (2017). Applying the bundle-move connection approach to the development of an online writing support tool for research articles. *Language Learning*, 67(4), 885-921. <https://doi.org/10.1111/lang.12250>
- Scott, M. (2019). *WordSmith Tools (Version 7)*. Stroud: Lexical Analysis Software.
- Schnur, E. & Csomay, E. (2020). Triangulating text segmentation methods with diverse analytical approaches to analyzing text structure. In J. Egbert & P. Baker (Eds.), *Using corpus methods to triangulate linguistic analysis*. (pp. 22-51). London/Yew York: Routledge. <https://doi.org/10.4324/9781315112466-2>
- Stewart, D. (2010). *Semantic prosody: A critical evaluation*. London: Routledge.
<https://doi.org/10.4324/9780203870075>
- Stubbs, M. (2008). Three concepts of keywords. Revised version of paper presented at Keyness in Text, Certosa di Pontignano, University of Siena, 2007, <http://www.uni-trier.de/fileadmin/fb2/ANG/Linguistik/Stubbs/stubbs-2008-keywords.pdf>
- Whitsitt, S. (2005). A critique of the concept of semantic prosody. *International Journal of Corpus Linguistics*, 10, 283-305. <https://doi.org/10.1075/ijcl.10.3.01whi>