Quantitative metrics to the CARS model in academic discourse in biology introductions

Charles Lam
Language Centre
School of Languages, Cultures and Societies
University of Leeds
Woodhouse Lane Leeds LS2 9JT
C.Lam@leeds.ac.uk

Nonso Nnamoko
Department of Computer Science
Edge Hill University
St Helens Road, Ormskirk L39 4QP
nnamokon@edgehill.ac.uk

Abstract

Writing research articles is crucial in any academic's development and is thus an important component of the academic discourse. The *Introduction* section is often seen as a difficult task within the research article genre. This study presents two metrics of rhetorical moves in academic writing: step-n-grams and lengths of steps. While scholars agree that expert writers follow the general pattern described in the CARS model (Swales, 1990), this study complements previous studies with empirical quantitative data that highlight how writers progress from one rhetorical function to another in practice, based on 50 recent papers by expert writers. The discussion shows the significance of the results in relation to writing instructors and data-driven learning.

1 Introduction

The research article is one of the most, if not the single most, important genres in academic discourse. The *Introduction* section in the research article is often reported to be difficult to write (Flowerdew, 1999; Hsu and Kuo, 2009).

Scholars have long recognized the central role of rhetorical moves in academic writing. The widely known analysis of the structure, the "Create a Research Space" (CARS) model (Swales, 1990, 2004) is the *de facto* standard in genre studies in academic discourse, alongside with the metadiscourse model by Hyland (2005, 2018). Swales (1990)'s CARS model observes the common pattern found in academic research articles, which encompasses three rhetorical moves (that can be seen as any textual unit, often one or more sentences, that aims to fulfill a particular function for a text). Each move can be decomposed to finer steps, while some steps are "optional", and some "obligatory" or expected. In the teaching setting, these moves and steps can be used to guide novice authors in presenting the context, purpose, objectives, literature review, and overall significance of their research logically and persuasively. The moves and associated steps (in bracket) are *Establishing a Territory* (define the field, provide background information, set the context), *Establishing a Niche* (identify a gap, problem, or unanswered question), *Occupying the Niche* (clearly state the purpose, focus, and objectives), *Reviewing Previous Research* (summarize relevant literature, critically review existing research), and *Establishing the Significance of the Research* (demonstrate the importance within the broader context).

A prevalent strand of studies under this tradition focuses on the correlation between particular rhetorical moves (e.g. *Establishing a Niche* or *Occupying the Niche*) and linguistic forms (e.g. frequent words or formulaic language, such as the n-gram "the aim of the study"). Beyond the study of lexical bundles, scholars often investigate the organizational structure of various parts of research articles from a qualitative perspective, while using empirical corpus data. Our study focuses on the structure of the *Introduction* section from an annotated corpus of biology research articles written by expert writers. While previous studies have investigated the same phenomenon, few works investigate the co-occurrence or relation between moves and steps at scale. For example, Samraj (2002, 2005) adopts a qualitative and manual close reading method with a few texts) for biology texts. In some cases, the focus is on the implementation of moves in actual linguistic forms (Lu et al., 2021, 2020), and the dataset were not made publicly available to facilitate follow-up studies or replication. As such, there is no existing dataset with clear annotation of the rhetorical moves.

This study presents our analysis of a small dataset of 50 texts in biology as a proof-of-concept and proposes two quantitative metrics to conduct move-step analysis. The contribution of this paper is two-fold: First, we discuss quantitative mea-

Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024), pages 71–77
March 21, 2024 ©2024 Association for Computational Linguistics

sures that allow for genre and rhetorical analysis without close reading by researchers, which is time-consuming and requires expert knowledge of genre analysis. Second, we outline our efforts in making the materials useful for writing instructors and novice learners of academic writing in higher education environments.

2 Related Work

Using corpus data to facilitate understanding of academic discourse is no novel approach. Specific to the English for Academic Purposes (EAP) community[1], there has been many corpora like the British Academic Written English (BAWE) corpus (Nesi and Gardner, 2018), Michigan Corpus of Spoken Academic English (MICASE) (Simpson et al., 2002), and the Michigan Corpus of Upper-Level Student Papers (MICUSP) (Römer and Swales, 2010). These resources have been widely used in the EAP community for analyzing academic language to facilitate materials development and instructions. The wider coverage of various disciplines means that the data are discipline-agnostic and capable of showing the overall patterns in the language of academic discourse.

To better understand rhetorical strategies through the CARS model, scholars have also employed corpus tools to investigate the use of common phrases associated with specific rhetorical moves. For example, combinations like "in this paper we present" and "it is well known that" are often found in the *Introduction* (Louvigné et al., 2014). Similarly, Jalali and Moini (2014) identify 161 common lexical bundles (i.e. frequent combinations of lexical items) in the *Introduction*. The most frequent ones in their study are often related to stating the purpose of the study, such as "The aim of the", "The objective of this", "study was to evaluate". Pérez-Llantada (2014) compares the skills in native and non-native speakers' of using formulaic combinations, using similar methods. While these findings provide solid evidence from attested examples used by writers, they are also limited in not addressing the organization of the *Introduction*, which is reported to be a common issue (Flowerdew, 1999).

Focusing on the organization and sequencing of the steps, scholars have also investigated how closely writers actually follow the CARS model in their practice. Previous studies have suggested that expert writers do not follow strictly the CARS model in their *Introductions* (Anthony, 1999; Samraj, 2002). Meanwhile, articles from different disciplines may display variations, e.g. applied linguistics (Ozturk, 2007), computer science (Orr, 1999; Maher and Milligan, 2019), engineering (Kanoksilapatham, 2015), and mathematics (McGrath and Kuteeva, 2012; Kuteeva and McGrath, 2015). Samraj (2005) discusses how introductions and abstracts of Wildlife Behavior and Conservation Biology, two closely related branches of biology, also show deviations from Swales' CARS model. Similarly, Milagros del Saz Rubio (2011) suggests that there are particular step-combinational patterns used (i.e. how rhetorical steps are assembled together) for achieving a variety of communicative purposes in agriculture.

3 Method

A total of 50 manuscripts from BioRxiv[2] were downloaded. From each of the five categories (Animal Behavior & Cognition, Biochemistry, Biophysics, Ecology, and Physiology), ten papers were randomly selected and annotated by the researcher.

The annotation is based on the original model by Swales (1990)[3], which includes three 'moves' essential to the introductory text, which can be further broken down into steps or options. In this study, each sentence is annotated with step label. The details are listed in Table 1. For simplicity, Moves are coded with 1-3, and Steps are coded with a-d, e.g. "Move 2 Step 3" is coded "2b".

4 Results

Taken the *introductions* of all the 50 articles together, the annotated small corpus contains 43,187 words and 1,297 sentences in total. Each category is represented by *introductions* of 10 articles. Table 2 shows the relevant statistics.

Figure 1 shows that Move 1 Step 3 'Reviewing previous research' is the most common type of

[1]The GENIA corpus (Kim et al., 2003), for example, was not designed for the purpose academic writing research or instruction. Rather, it was designed for knowledge mining in biology.

[2]https://www.biorxiv.org/

[3]While a revised model is proposed in Swales (2004) with the aim to better accommodate variations in response to some critiques (see e.g. Anthony (1999); Samraj (2002); Ozturk (2007)), the updates (e.g. grouping all steps in Move 1 to "Topic generalizations of increasing specificity" (Swales, 2004, 230) do not appear to generate concrete steps that can better account for variations. Rather, the updated description accommodates a wider range of variations simply by being more generic. For the practical purpose of annotation, this study uses the original scheme with more fine-grained steps.

Table 1: Steps in the CARS model (Swales, 1990)

Move/Step	Description	Code
Move 1	*Establish Research Territory*	
Step 1	Claiming centrality	1a
Step 2	Making topic generalizations	1b
Step 3	Reviewing previous research	1c
Move 2	*Establish a Niche*	
Option 1	Counter-claiming	2a
Option 2	Indicating a gap	2b
Option 3	Question-raising	2c
Option 4	Continuing a tradition	2d
Move 3	*Occupy the Niche*	
Step 1a	Outlining purposes	3a
Step 1b	Announcing present research	3b
Step 2	Announcing principal findings	3c
Step 3	Indicating article structure	3d

Table 2: Mean word counts and sentence counts per file

Category	Mean Word Count	Mean Sentence Count
Animal Behv & Cogn	714.8	20.4
Biochemistry	836.8	27.8
Biophysics	883.1	28.4
Ecology	1077.9	26.7
Physiology	806.1	26.4

sentence in the data.

4.1 Step Collocation

To better understand the sequencing of rhetorical steps, this study proposes a simple measure of step-*n*-grams that captures the common sequences of steps. In the data, the same steps tend to span over multiple sentences, which likely signals the same rhetorical function expressed by multiple sentences. For example, the segment[4] in Table 3 was coded as 1b-1c-2b in step-*n*-gram, where the repetition of 1c over three sentences is coded as one single step.

Excluding these repetition of the same steps, there are 169 attested combinations. The most common step-*n*-grams are listed in Table 4:

[4] https://doi.org/10.1101/2023.10.29.564363
The biology texts are not cited in this study as they are used as textual data, not academic citation.

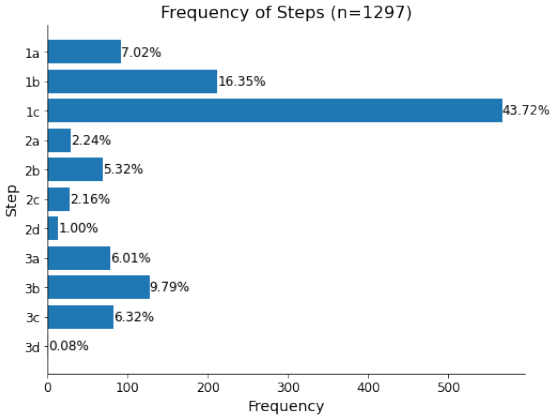

Figure 1: Frequency of steps (n=1,297)

The results in Table 4 indicate that the rhetorical progression (i.e. moving from one step to another) "1a-1b-1c" is common, occurring in 34 out of the 50 texts. For bigrams "1b-1c" (n=62) and "1a-1b" (n=51), we even observe repetition within the texts, as their frequencies are higher than the number of texts (n=50). It is not surprising that the step 1c occurs in almost all combinations, due to its central role to review previous studies and thus the high frequency. The second highest step-3-gram is "1b-1c-2b" (n=18), which can also be explained by the high frequency of step 2b "Indicating a gap", and how it connects the steps "Making topic generalizations" and "Indicating a gap", which is the most frequent option among the four in Move 2. See more in section 4.3.

4.2 Lengths of Steps

The length of step measures how many sentences the same step may span over in a contiguous manner. Table 5 shows the lengths of all the steps. Values of 0 indicate that the step can be absent in some texts. Step "1c - Reviewing previous research" is the only step that is never skipped in the attested data. The step is also the longest among all steps. Again, this is not surprising given its central role.

On the other hand, most other steps are much shorter, as indicated by their maximum lengths and mean lengths. The discussion will further defend the use of this seemingly mundane information from a pedagogical perspective for students or even novice writers.

4.3 How to "Establish a Niche" (Move 2)

The classic CARS model includes four options or approaches to implement the rhetorical move of establishing a niche. That is, scholars decide whether

Table 3: A multi-sentence step in "1b-1c-2b"

Step & Sentence
[1b]: Cancer cells grow in a microenvironment wherein they closely interact with the extracellular matrix (ECM).
[1c]: As a major ECM component, collagen composition regulates various steps of cancer progression including growth, invasion, and metastasis, partly through activation of its canonical receptor integrin to regulate cytoskeleton organization and cell motility [5–7].
[1c]: Recently, discoidin domain receptor tyrosine kinase 2 (DDR2), a non-typical collagen receptor that is dysregulated in various cancer types, has emerged as a key signaling molecule in carcinogenesis [8, 9].
[1c]: Collagen binding to DDR2 activates its tyrosine kinase activity to initiate canonical pathways such as ERK/MAPK and PI3K/AKT signaling cascades [10–12].
[2b]: Despite these studies, how DDR2 regulates cancer cell behavior is incompletely understood.

Table 4: Top 5 step-bigrams and step-trigrams

Step-Bigram	Freq	Step-Trigram	Freq
1b-1c	62	1a-1b-1c	34
1a-1b	51	1b-1c-2b	18
1c-2b	38	1b-1c-1b	17
1c-1b	29	1c-1b-1c	17
2b-1c	23	1c-2b-1c	15

they are making a counter-claim (e.g. "However, this validity may not be related to the neurobiology of depression"[5]) or to indicate a research gap (e.g. "Despite these studies, how DDR2 regulates cancer cell behavior is incompletely understood."[6]) in order to show the niche of their own study. It has been made clear that these options are not mutually exclusive, nor do they follow any particular hierarchy or ordering. Authors from our data often adopts the option of "Indicating a gap". Almost half of the 139 examples of Move 2 are from option 2 (Option 1 = 20.86%, n=29, Option 2 = 49.64%, n=69, Option 3 = 20.14%, n=28, Option 4 = 9.35%, n=13). It is, however, important to note that these options are not mutually exclusive. The same introduction may contain multiple options by both indicating a gap (option 2) and raising a question (option 3).

[5] https://doi.org/10.1101/2023.11.08.566266

[6] https://doi.org/10.1101/2023.11.03.565457

Table 5: Lengths of steps

Step	Min	Max	Mean
1a	0	4	1.28
1b	0	10	2.14
1c	1	16	3.97
2a	0	5	1.61
2b	0	6	1.33
2c	0	4	1.27
2d	0	3	1.44
3a	0	6	1.47
3b	0	8	1.98
3c	0	7	2.65
3d	0	1	1

5 Discussion

In the EAP community, studies on rhetorical moves are abundant, especially with the focus on the correlation between lexical bundles and particular rhetorical moves, i.e. what phrases appear in which moves/steps (Cortes, 2013; Staples et al., 2013; Moreno and Swales, 2018; Omidian et al., 2018; Appel, 2022). To complement this strand of research that focuses on language use, the present study discusses the progression of the moves and steps. By introducing quantitative measures, we have identified the distribution of specific steps, as well as how different steps may collocate with each other. Potentially, a scaled up version using similar methods will be able to identify any micro-variations across sub-disciplines, as some previous studies suggest.

Our results also confirms what Samraj (2005) argues with regard to the deviations from the classic CARS model. In our sentence-by-sentence annotation, it is often found that Move-1 Step-3 ("Reviewing previous research") is interspersed with other moves. It can be explained by the need to provide further support from previous studies, once the authors have made topic generalizations (see bigram "1b-1c": n=62) or indicated a gap (see bigram "2b-1c": n=23).

While the quantitative results from the step-*n*-gram and lengths of steps may seem mundane, novice scientific writers can use these numerical results as quick reference. The attested data in the annotated corpus will also facilitate material development. Rather than prescribing to students[7]

[7] In the authors' context, the students are all at the post-

that the *Introduction* must follow a certain pattern, students can see both conformity to and deviation from the standard CARS model. This allows students to gain better understanding of how expert writers may consciously depart from the CARS model.

Given the internationalization of many institutions and the increasing needs for support in academic literacy to both students and early career researchers, the findings here may also mean that instructions to discipline-specific writing should be more fine-grained. For instance, students in biodiversity would have different needs and writing models from students in molecular biology. Annotated corpus data will allow instructors to easily find attested data for various needs of students.

6 Conclusion and Future Work

This study has shown results from a small annotated corpus and how they enhance our understanding of academic discourse through the lens of the CARS model. The study bears implications on our understanding of progression in rhetorical across steps (through step collocation) and implementation of steps (through lengths of steps), which in turn benefits teaching of academic writing. In future research, it may also be interesting to investigate whether there is any significant differences between preprints (e.g. from BioRxiv as in the present study) and published research articles. While both kinds of data are supposed to be written by advanced or expert writers, there appears to be little research on the contribution of peer review and editing specific to the rhetorical quality of the articles. We acknowledge that the dataset is limited by its size and the single annotator, and intend to remedy these limitations in our ongoing work.

In future work, we aim to enhance the efficiency of the annotation process through the application of semi-supervised learning techniques. This involves leveraging the manually annotated corpus to develop an enriched corpus. For example, training a KNN model will be useful for the multi-class task that classify the sentences into the various steps. Additionally, we can also implement few-shot learning methodologies with the moves and steps being vectorised with pre-trained LLMs, such as GPT (Brown et al., 2020), on the modest "labelled" dataset to develop machine learning models that can generalise and make accurate classifications on new data samples.

graduate level of MSc in biology programs, with a mix of L1 and L2 users of English.

References

Laurence Anthony. 1999. Writing research article introductions in software engineering: How accurate is a standard model? *IEEE transactions on Professional Communication*, 42(1):38–46.

Randy Appel. 2022. Lexical bundles in l2 english academic texts: relationships with holistic assessments of writing quality. *System*, 110:102899.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Viviana Cortes. 2013. The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for academic purposes*, 12(1):33–43.

John Flowerdew. 1999. Problems in writing for scholarly publication in english: The case of hong kong. *Journal of Second Language Writing*, 8(3):243–264.

Yu-kai Hsu and Chih-Hua Kuo. 2009. Writing RA introduction: Difficulties and strategies. In *2nd International Conference on English, Discourse, and Intercultural Communication, Macau, China*. Citeseer.

Ken Hyland. 2005. *Metadiscourse*. London: Continuum.

Ken Hyland. 2018. *Metadiscourse: Exploring interaction in writing*. Bloomsbury Publishing.

Zahra Sadat Jalali and M Raouf Moini. 2014. Structure of lexical bundles in introduction section of medical research articles. *Procedia - Social and Behavioral Sciences*, 98:719–726.

Budsaba Kanoksilapatham. 2015. Distinguishing textual features characterizing structural variation in research articles across three engineering subdiscipline corpora. *English for Specific Purposes*, 37:74–86.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Maria Kuteeva and Lisa McGrath. 2015. The theoretical research article as a reflection of disciplinary practices: The case of pure mathematics. *Applied Linguistics*, 36(2):215–235.

Sébastien Louvigné, Jie Shi, and Sonia Sharmin. 2014. A corpus-based analysis of the scientific RA genre and RA introduction. In *Proceedings of the 2014 International Conference on Advanced Mechatronic Systems*, pages 123–127.

Xiaofei Lu, J Elliott Casal, and Yingying Liu. 2020. The rhetorical functions of syntactically complex sentences in social science research article introductions. *Journal of English for Academic Purposes*, 44:100832.

Xiaofei Lu, Jungwan Yoon, Olesya Kisselev, J. Elliott Casal, Yingying Liu, Jinlei Deng, and Rui Nie. 2021. Rhetorical and phraseological features of research article introductions: Variation among five social science disciplines. *System*, 100:102543.

Paschal Maher and Simon Milligan. 2019. Teaching master thesis writing to engineers: Insights from corpus and genre analysis of introductions. *English for specific purposes*, 55:40–55.

Lisa McGrath and Maria Kuteeva. 2012. Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. *English for Specific Purposes*, 31(3):161–173.

M. Milagros del Saz Rubio. 2011. A pragmatic approach to the macro-structure and metadiscoursal features of research article introductions in the field of agricultural sciences. *English for Specific Purposes*, 30(4):258–271.

Ana I Moreno and John M Swales. 2018. Strengthening move analysis methodology towards bridging the function-form gap. *English for specific purposes*, 50:40–63.

Hilary Nesi and Sheena Gardner. 2018. The BAWE corpus and genre families classification of assessed student writing. *Assessing Writing*, 38:51–55.

Taha Omidian, Hesamoddin Shahriari, and Anna Siyanova-Chanturia. 2018. A cross-disciplinary investigation of multi-word expressions in the moves of research article abstracts. *Journal of English for academic purposes*, 36:1–14.

Thomas Orr. 1999. Genre in the field of computer science and computer engineering. *IEEE Transactions on Professional Communication*, 42(1):32–37.

Ismet Ozturk. 2007. The textual organisation of research article introductions in applied linguistics: Variability within a single discipline. *English for Specific Purposes*, 26(1):25–38.

Carmen Pérez-Llantada. 2014. Formulaic language in l1 and l2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes*, 14:84–94.

Ute Römer and John M Swales. 2010. The Michigan corpus of upper-level student papers (MICUSP). *Journal of English for Academic Purposes*, 9(3):249.

Betty Samraj. 2002. Introductions in research articles: Variations across disciplines. *English for specific purposes*, 21(1):1–17.

Betty Samraj. 2005. An exploration of a genre set: Research article abstracts and introductions in two disciplines. *English for specific purposes*, 24(2):141–156.

Rita Simpson, Sarah Briggs, Janine Ovens, and John M Swales. 2002. The Michigan corpus of upper-level student papers (MICUSP). http://quod.lib.umich.edu/cgi/c/corpus/corpus. Accessed: 2023-11-30.

Shelley Staples, Jesse Egbert, Douglas Biber, and Alyson McClair. 2013. Formulaic sequences and eap writing development: Lexical bundles in the toefl ibt writing section. *Journal of English for academic purposes*, 12(3):214–225.

John M Swales. 1990. *Genre analysis: English in academic and research settings*. Cambridge University press.

John M Swales. 2004. *Research genres: Explorations and applications*. Cambridge University Press.

A Appendix

Figure 2 shows the sequence of the steps in all the 50 annotated texts.

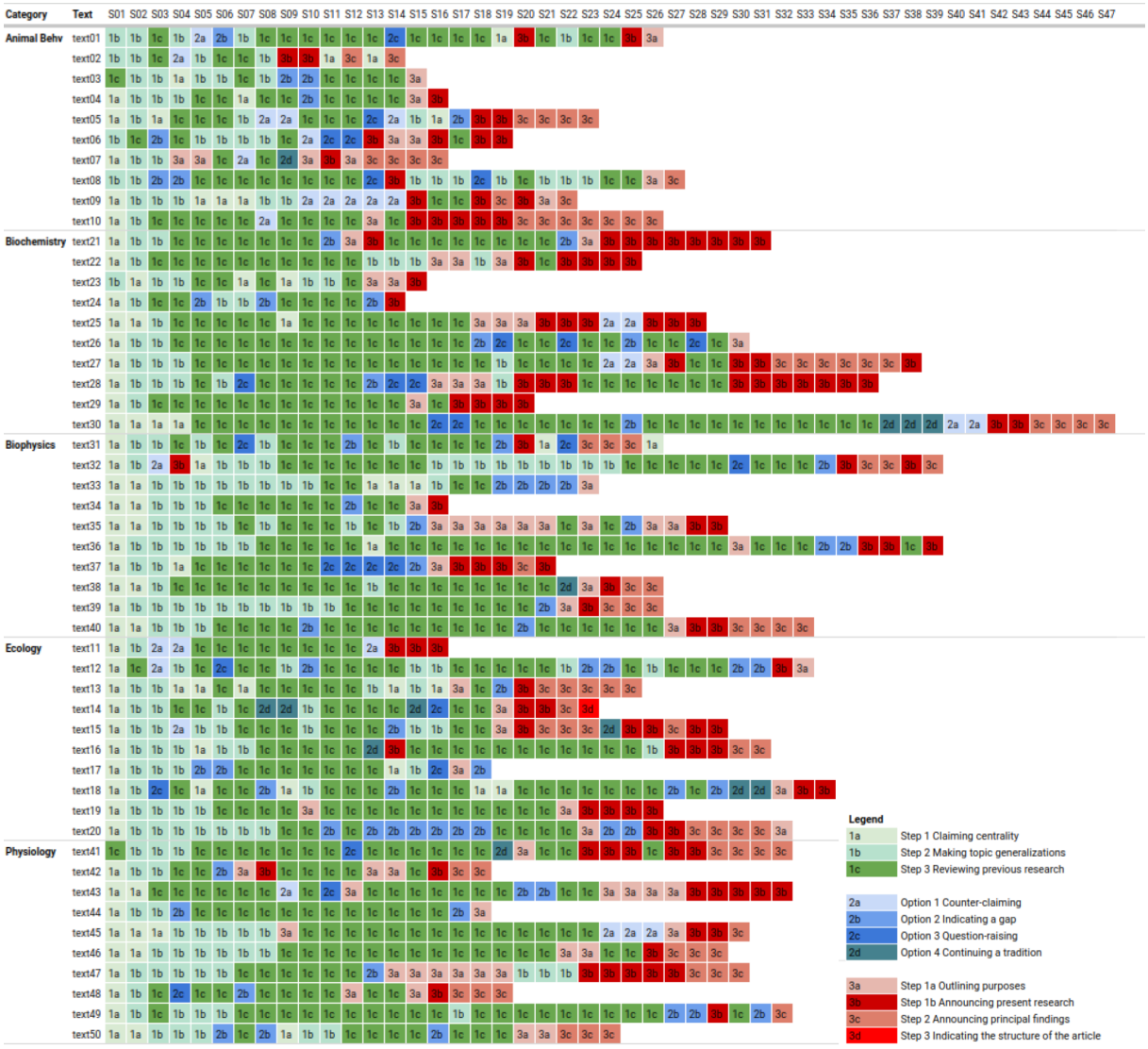

Figure 2: Sequencing of the Steps