









RESEARCH ARTICLE

An analysis of form and function of a research article between and within publishers and journals

Sarah Nathan¹ , Leah Haynes¹ , Jessica Meyer¹ , Josh Sumner¹ ,
Cynthia Hudson-Vitale^{1,2} , and Leslie D. McIntosh¹ 

¹Ripeta LLC, St. Louis, MO

²Association of Research Libraries

an open access  journal



Keywords: article headings, information organization, machine learning, scholarly communication, science communications

ABSTRACT

The identification and subsequent analysis of research articles for machine learning and natural language processing is a complicated task given the lack of consistent article organization principles and heading naming conventions across publishers and journals. Given this, an understanding of how research articles organizationally follow a common function and their use of various heading terms, or forms, is a critical step in applying machine learning techniques for data and information mining across a corpus of articles. To address this need, the authors developed and implemented an article heading form and function analysis across 12 publishers including both research articles and nonresearch articles. Our aim was to (a) identify each of the labeled sections used by research articles, define these sections based on their rhetorical function, and determine frequency of use; (b) within the given data set, determine all of the alternative labels used to identify these sections; and (c) determine whether these sections can be used to consistently determine (1) whether an article is a true research article, or (2) whether an article is *not* a research article. The results indicated wide variability in the organization of research articles with 24 common sections, known by 186 different names both within and across publishing houses.

1. INTRODUCTION

From biology to architecture to writing, a simple principle holds true: Form must align with function. In scientific writing, the empirical research article (RA) is the form used to communicate new, systematically tested ideas in a way that allows those ideas to be evaluated by other experts (Tanti, 2014). As a mechanism for scholarly communications, the form and structure of the RA has been the study of much research and analysis, specifically in the areas of genre classification, or successfully differentiating texts that belong to different genres, and communicative moves, or the mapping of rhetorical structure within an article section (Swales, 1981).

There is also a growing body of research on the structure of an RA, specifically around article headings. Thelwall (2019) leveraged a large corpus of full text articles from PubMed Central to look across domains to compare the structure of RA headings. Leveraging the high-quality meta-data within PubMed Central, Thelwall concluded that there was very little consistency in the structure of a research article both within and across the many scientific disciplines. Thelwall's work built on the research of Teufel (1999), who determined that scientific texts could

Citation: Nathan, S., Haynes, L., Meyer, J., Sumner, J., Hudson-Vitale, C., & McIntosh, L. D. (2021). An analysis of form and function of a research article between and within publishers and journals. *Quantitative Science Studies*, 2(2), 643–661. https://doi.org/10.1162/qss_a_00135

DOI:
https://doi.org/10.1162/qss_a_00135

Peer Review:
https://publons.com/publon/10.1162/qss_a_00135

Received: 31 August 2020
Accepted: 27 April 2021

Corresponding Author:
Cynthia Hudson-Vitale
cvitale@arl.org

Handling Editor:
Ludo Waltman

Copyright: © 2021 Sarah Nathan, Leah Haynes, Jessica Meyer, Josh Sumner, Cynthia Hudson-Vitale, and Leslie D. McIntosh. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.



be arranged in key argumentative zones based on section function and expected moves. Additional research has focused on mapping the research article structure and form of specific disciplines and subdisciplines. For example, Kanoksilapatham (2015) found variations and unique characteristics for each engineering subdiscipline RA structure. Similarly, Tessuto (2015), analyzed the form and moves of empirical law RAs, finding that the typical Introduction, Methods, Results, and Discussion (IMRD) framework for organizing an RA was not found in contemporary outputs published by the law discipline.

What hasn't been widely studied within this area of critical analysis is the form and function of the research article between and within journals and publishers. While some publishers, journals, and scholarly societies have expectations and guidelines for reporting research or a framework for RA structure (ICMJE, 2004; *Journal of Environmental Quality*, 2021), no known analysis has completed a comparison across journals within a group of publishers. To conduct this research, the aims are to

1. Identify each of the labeled sections used by RAs, define these sections based on their rhetorical purpose, and determine frequency of use.
2. Within the given data set, determine all of the alternative labels used to identify these sections.
3. Determine whether these sections can be used to consistently determine (a) whether an article is an RA, or (b) whether an article is *not* an RA.

An article "section" is defined as any segment of text that has been set apart from the main text with a label. A "subsection" is any segment of text that falls within a section and has been set apart from that section with its own label. A "label" is a word or phrase used to identify and describe a section or subsection.

Given that this research is bound by publisher/journal and not by academic discipline, no large-scale corpus, with CC-BY or full-text open access, such as PubMed Central is available to identify and parse RAs and the structured metadata. Thus, the aims require a process for first doing this. The first aim seeks to catalogue and define the current RA structure, identifying rhetorical elements and cataloguing how frequently they are used to determine which components are obligatory in an RA and which are optional. By defining an RA and its specific elements, it is possible to determine how to differentiate RAs from other similar genres.

The second aim seeks to document variation in how sections are labeled, thus identifying the particular problems that machine programmers face when identifying RAs and locating information within RAs.

Finally, our third aim seeks to compare sections and labels used in RAs across articles within specific publishers and journals to those used by review articles, meta-analyses, and case studies to determine if there are already straightforward ways to differentiate these genres based on article structure.

1.1. Literature Review

In the last two decades, a considerable body of work has emerged to document the rhetorical components of RAs. This research has documented the macrosections of the RA, the various orders in which those sections can appear, and the specific rhetorical functions of those sections. Together, these existing studies have built a rich framework for understanding the structure and definition of RAs. Yet the dichotomous focus on either macrostructure or the rhetorical

components of a particular section has left midlevel structure relatively overlooked. Given that RAs, non-RAs, and nonresearch articles often contain the same macrosections, RA subsections within publishers and specific journals deserve focus.

The current literature builds generally from Swales' (1981) seminal work on RA structure, in which he established move-step analysis. In the Swalesian tradition, a move can be defined as "a text segment that performs a communicative function, contributing to the global function of a whole text. Moves can vary in length ... and can be recognized by a set of linguistic features" (Kanoksilapatham, 2015). In turn, moves are broken down into steps, essentially rhetorical subunits that work together to accomplish a move. For example, Dobakhti (2013) identifies one RA move, "Commenting on Findings," as composed of three steps: explaining, interpreting, and evaluating.

Authors use these moves to identify and define the macrosections of RAs. For instance, Lin and Evans (2012) use move-step analysis to question the breakdown of RAs into only the four main sections of IMRD. They argue that the Literature Review and Conclusion sections contain moves not accounted for in prior move-step analyses of I, M, R, and D and should, therefore, be their own sections. Lin and Evans (2012) also posit that, when the Results and Discussion sections are combined into a single section, that section contains a different set of moves than those in separate R and D sections. Thus, according to Lin and Evans (2012), separate Results and Discussion sections are definitionally distinct from a combined Results and Discussion section. Authors continue to disagree on how these sections should be categorized, and many authors continue to lean on the IMRD breakdown despite Lin and Evans' (2012) analysis. For instance, Li and Ge (2009) use Nwogu's (1997) set of moves and IMRD framework to conduct their analyses of articles published in 1985 and 2004. Kanoksilapatham (2015) also focuses on the I, M, R, and D sections when analyzing how moves vary across engineering papers, and Hsieh, Tsai, Lin, Luoi, and Kuo (2006) treat the Conclusion as a variation of the Discussion section.

At an even more fundamental level, researchers have used linguistic patterns to identify and define the moves themselves. For example, de Waard and Pander Maat (2012) and Dahl (2009) use verb tense patterns to define the rhetorical purpose of particular sentences or paragraphs. By contrast, Kashiha (2015) identifies phrases or "lexical bundles" that are used to introduce particular rhetorical components. These lexical bundles contain clues about the function of the subsequent text, allowing Kashiha (2015) to define that text's purpose and categorize it as a move.

2. METHODS

2.1. Sample Selection and Inclusion Criteria

RAs published from 2007–2020 were collected from the journals of ten major publishers: Cambridge University Press, DeGruyter, Emerald, IOP Publishing, Karger, Oxford University Press, PLOS ONE, Sage, Taylor & Francis, and Wiley. Within each publisher, five to seven journals were sampled. Papers were selected randomly from across all disciplines, and only open access articles were considered. Five articles from each journal were analyzed, except when a journal did not have enough open access articles.

Case studies, meta-analyses, and review articles from journals of 12 major publishers were harvested: Cambridge University Press, DeGruyter, Elsevier, Emerald, Karger, Nature, Oxford University Press, PLOS ONE, Sage, Science, Taylor & Francis, and Wiley. More publishers were included in this harvest because of the difficulty of finding sufficient open-access case studies, meta-analyses, and review articles.

2.2. Data Collection

The goal of the data collection process was to record all of the labeled sections and subsections within RAs, meta-analyses, case studies, and review articles, as well as all of the different labels used to identify those sections. Before reviewing actual articles, one author (LH) developed a preliminary list of sections and labels based on previous work annotating thousands of RAs both by hand and with the use of Prodigy (2017). Prodigy is an annotation tool from the creators of spaCy (2017) that produces training and evaluation data to develop machine learning models.

Using this list of preliminary sections, JM conducted the initial textual analysis, manually scanning each article and recording when one of the anticipated sections was present. When she encountered a section with a different rhetorical purpose from all previously recorded sections, JM created a new column in the spreadsheet and began recording that section for all subsequent papers. When the found section label was a verbatim match for the label in the spreadsheet, JM simply tallied the result. If JM encountered a section that she suspected was an alternate label for one of the pre-established sections, she recorded the verbatim wording in the spreadsheet under that section's column. After scanning and recording all sections in a given article, JM then re-examined the article to annotate which sections were main sections and which were subsections. Sections that were demarcated with an individualized label (i.e., the label was specific to the topic of the paper) were not included in the tally. To review the full data collection workflow, see Figure 1.

2.3. Qualitative Data Analysis: Delineating Sections and Crafting Definitions

After the initial data collection, a multistep process was developed to delineate our final list of sections and to craft definitions for these sections (see Figure 2). First, the research team reviewed all the section labels and noted any suspected overlap in rhetorical purpose among sections with different names. For each overlapped grouping, it was first determined whether more than one of the sections ever appeared in a single article. If so, these sections were kept separate because they served different purposes. If not, the article texts were reviewed to determine the overall purpose of each section; any notable differences in specific language among these sections; and words or phrases held in common by these sections. If both the purpose and the language aligned, these

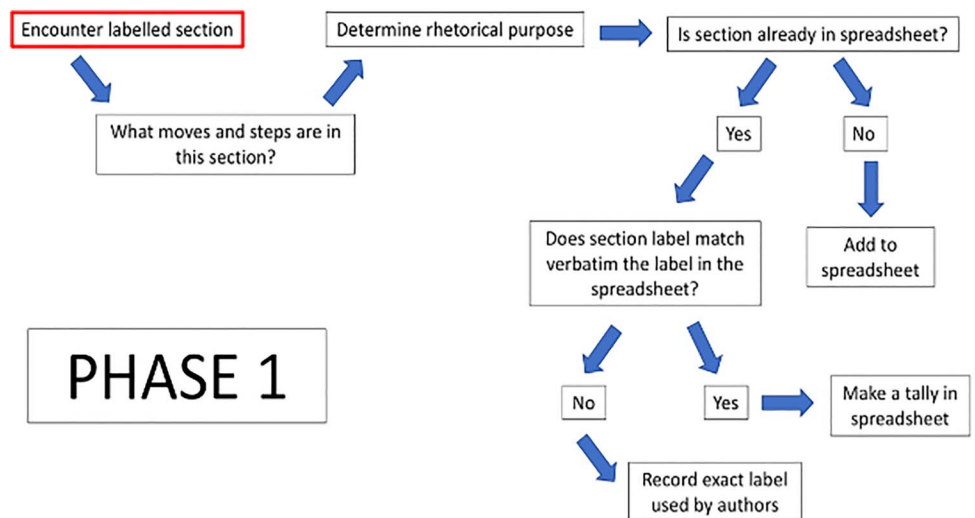


Figure 1. Initial data collection workflow.

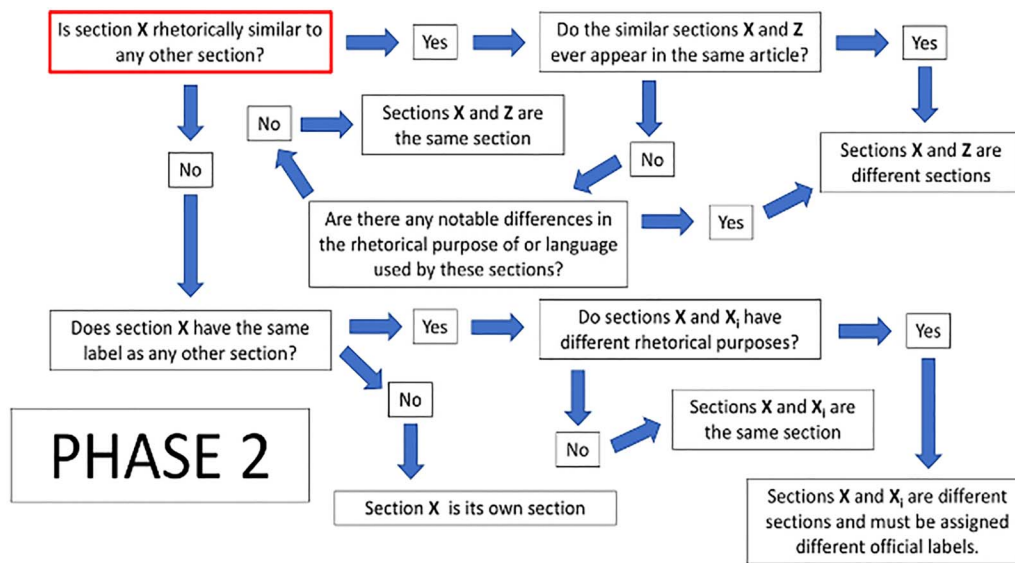


Figure 2. Workflow to determine whether or not sections with similar purposes or identical labels were distinct sections.

sections were provisionally combined. In the final step, journal author guidelines, RA writing guides, and previous move-step analyses that discussed these specific sections were reviewed. These texts were used to refine the purpose assigned to each section and to identify any stated differences in purpose for grouped sections.

An inverse process was used to determine when the same label was used to refer to more than one type of section (see Figure 2). Annotators recorded when they suspected that one verbatim label was used to refer to sections with different purposes. They additionally identified the pre-established section with which each verbatim label most aligned. The articles' texts were then reviewed to verify that sections with the same verbatim label did, in fact, have divergent purposes, and determine if these labels could be added to an overlapping group, as described in the previous paragraph.

With the finalized list of sections, each section was assigned a title based on the labels most commonly used by authors. SN then crafted definitions in a three-step process (see Figure 3). She

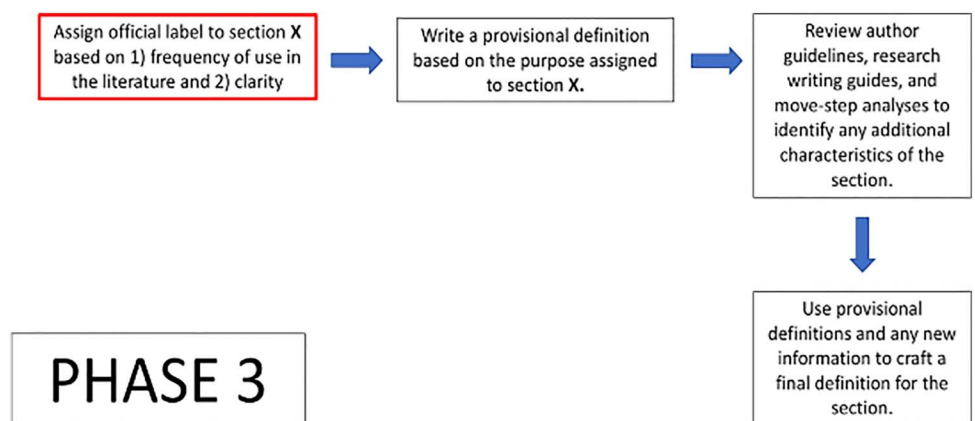


Figure 3. Workflow for crafting section definitions.

<p>RESEARCH ARTICLE ARTICLE TITLE: SAMPLE</p> <p> OPEN ACCESS</p> <p>Author Name-one*, Author Name-two*, Author Name-three* *Associated institution, *Associated institution, *Associated institution *Corresponding Author</p> <p>Keywords: _____, _____, _____</p> <p>ABSTRACT _____ _____ _____ _____</p> <p>INTRODUCTION _____ _____</p> <p>OBJECTIVES: goal(s) of research and/or research question _____</p> <p>BACKGROUND _____ _____ _____</p> <p>METHODS _____ _____ _____ _____</p> <p><i>Ethics Approval:</i> indication of whether researchers obtained approval from an institutional review board _____</p> <p>MATERIALS _____ _____</p> <p>RESULTS _____ Statistical Analysis: _____ _____ _____</p>	<p>DISCUSSION _____ _____ _____ _____</p> <p>Limitations: _____ _____ _____</p> <p>CONCLUSION _____ _____</p> <p>ACKNOWLEDGEMENTS _____</p> <p><i>Author Contributions:</i> details about authors' roles in developing & publishing manuscript <i>Conflicts of Interest:</i> acknowledgement of relationships which may influence the research and findings <i>Funding:</i> whether authors received funding, role of funder(s), & specific grant information</p> <p>SUPPLEMENTARY INFORMATION _____ _____</p> <table style="width: 100%; border: none;"> <tr> <td style="width: 50%; vertical-align: top; padding: 5px;"> <p>DATA AVAILABILITY STATEMENT _____ _____ _____</p> <p>TRANSPARENCY STATEMENT declaration affirming that the manuscript is an honest and accurate account of the research, that no important aspects of the study have been omitted, & that any changes to the study were adequately explained</p> <p>REFERENCES publication information about all outside sources of information used to inform the research</p> </td> <td style="width: 50%; vertical-align: top; padding: 5px;"> <p>CODE AVAILABILITY STATEMENT _____ _____</p> <p>PUBLICATION DETAILS _____ _____ _____ _____</p> </td> </tr> </table>	<p>DATA AVAILABILITY STATEMENT _____ _____ _____</p> <p>TRANSPARENCY STATEMENT declaration affirming that the manuscript is an honest and accurate account of the research, that no important aspects of the study have been omitted, & that any changes to the study were adequately explained</p> <p>REFERENCES publication information about all outside sources of information used to inform the research</p>	<p>CODE AVAILABILITY STATEMENT _____ _____</p> <p>PUBLICATION DETAILS _____ _____ _____ _____</p>
<p>DATA AVAILABILITY STATEMENT _____ _____ _____</p> <p>TRANSPARENCY STATEMENT declaration affirming that the manuscript is an honest and accurate account of the research, that no important aspects of the study have been omitted, & that any changes to the study were adequately explained</p> <p>REFERENCES publication information about all outside sources of information used to inform the research</p>	<p>CODE AVAILABILITY STATEMENT _____ _____</p> <p>PUBLICATION DETAILS _____ _____ _____ _____</p>		

Figure 4. Graphic showing the most common format of a research article.

began by writing provisional definitions based on the purposes that were assigned to each section. She then reviewed author guidelines, RA writing guides, and move-step analyses to create a list of common features and purposes for each of the sections. Finally, she used our provisional definitions and the list of common features and purposes to create definitions that captured the scientific community's collective understanding for each of the sections.

The sections identified typically appear in a standard order. Based on the collected data and observations, section definitions, American Association for the Advancement of Science (AAAS) (n.d.) guidelines, and the generally accepted IMRD format for academic articles, a model for an RA was developed that follows established practices. This model is presented in Figure 4.

2.4. Data Validation

Once the list of section labels and their definitions was finalized, the research team conducted a set of validation checks. SN conducted this second set of checks, scanning each of the RAs, case studies, meta-analyses, and review articles. These checks consisted of two parts and ensured that no sections had been missed or had been mistakenly tallied when actually absent and that suspected alternate labels were correctly categorized. This second step required that SN skim all sections with alternate labels to ascertain their purpose and check that the purpose matched the definitions. In cases when the appropriate categorization was not clear, the researcher presented the found text to the larger group for final deliberation.

2.5. Statistical Analysis

To analyze for differences in section frequency between RA and nonresearch articles, a two-proportion z test and used a 95% confidence interval (CI) was conducted. This statistical proportion test allowed us to test our hypothesis and is appropriate as the data are approximately normally distributed, of suitable size, and independent (McCullagh & Nelder, 1989). Tests were run with R statistical software version (R Core Team, 2014) 3.5.1 (2018-07-02) and the Mosaic package (Pruim, Kaplan, & Horton, 2017) (mosaic_1.5.0).

3. RESULTS

3.1. Objectives 1 and 2: Section Labels, Definitions, and Article Structure

The first objective of this study was to identify each of the labeled sections used by RAs, define these sections based on their rhetorical purpose, and determine the general structure of an RA based on the frequency with which these sections are used. Our second objective was to identify all of the alternative labels used to identify these sections.

To achieve these objectives, 250 RAs and 30 non-RAs were analyzed (ten case studies, ten review articles, and ten meta-analyses). Within the RAs, 31 different section types known collectively by 302 different labels (see Tables 1 and 2) were identified. Twenty-four sections that are theoretically applicable to every RA, known by 186 different names (see Table 1) were found. The additional seven sections, all subsections of the “Methods” section, are only relevant to certain types of research, and were known by 116 different labels (See Table 2).

Some of the sections in Table 1 can appear as either a main section or as a subsection of another section. For example, Statistical Analysis almost always appears as a subsection of Methods or Results.

The research team also found that some specific labels were used to refer to more than one type of section. For example, “Summary” may refer to either the Abstract or Conclusion. Additionally, “Replication of Results,” “Availability of Data and Materials,” and “Open Practices Statement” may refer to Code Availability Statement, Data Availability Statement, or both combined. “Instruments” could also be used to refer either to the Instruments section or to the Materials section.

Many sections tend to be combined under one label, the most obvious examples being “Materials and Methods” and “Code Availability Statement and Data Availability Statement.” Combination is especially frequent among the subsections in the Methods section. For instance, Study Subjects and Selection Criteria are often combined, as are Study Design and Procedures. For the purposes of this study, the project team counted combined labels toward each of the individual sections. For example, if an article had a “Study Design and Procedures” section, a tally for both Study Design and Procedures was made, but it also was recorded that it was a combined label. Common examples of combined labels are listed in Table 3.

A second type of combination, which is much more difficult to detect, occurs when authors combine sections but only list one of the labels rather than both. Interestingly, this type of combination is most prevalent with the Conclusion, Discussion, and Results sections. For example, regarding the Results, Discussion, and Conclusion sections, PLOS ONE states (PLOS ONE, n.d.), “These sections may all be separate, or may be combined to create a mixed Results/Discussion section (commonly labeled ‘Results and Discussion’) or a mixed Discussion/Conclusions section (commonly labeled ‘Discussion’). These sections may be further divided into subsections, each with a concise subheading, as appropriate.” In these instances, the

Table 1. Twenty-four sections, known by 186 unique labels, were relevant to all research articles and could be defined by consistency of use

Section	Sample alternative labels	Definition
Abstract	Summary	<p>The abstract is a concise summary of the article or study that is able to stand on its own. It must describe the major aspects of the entire paper, including:</p> <ul style="list-style-type: none"> • A brief statement that provides background on the topic; • The overall purpose and the research problems that were investigated; • The basic design of the study; • Major findings; and • A brief summary of the author's interpretations and conclusions.
Introduction	N/A	The introduction describes the significance of the topic, establishes the gap in current knowledge in literature associated with the topic, often includes a literature review with background information relevant to the key questions, and outlines the objectives or hypothesis of the research.
Objectives*	Statement of Novelty; Statement of Purpose; Statement of Problem; Problem Statement; Hypothesis; Aim(s); Aim(s) of the Study; Aims and Significance of the Study; Research Question(s); Objectives, Scope, and Novelty	The objectives state the goal of the research or the question that the research will answer, often accompanied by a brief description of how the researcher will achieve those aims. The objectives should clearly relate to the gap in current research that the author establishes in the introduction.
Background*	Literature Review; Theoretical Framework; Previous Studies; Previous Work; Theory; Context	Typically a subsection of the introduction, the background provides a review of the literature associated with the specific research topic, describing the current state of knowledge about the issue, and exposing the information gap that the research will address.
Methods	Methods and Materials; Methodology; Methodological Section; Research Method/Methods/Methodology; Data and Methods/Methodology; Methods/Methodology and Data; Experimental Approach/Design; Method Summary; Full Methods; (Experimental) Procedure(s); Approach; Implementation; Physical/Experimental Setup; Methodological considerations; Patients and Methods; Study Site and Methods; Methods of Analysis; Analytical Methods; Method(s) and Measures; Method(s) of Data Collection; Methodology and Theories; Design and Methods; Empirical Framework and Data; Experimental Details; Research Design	<p>The methods section clearly describes the specific design of the study and provides a description of the procedures that were performed, giving enough detail that the reader can assess the credibility of the results. A methods section typically contains:</p> <ul style="list-style-type: none"> • The population and equipment used in the study; • How the population and equipment were prepared and what was done during the study; • The protocol used; • The outcomes and how they were measured; • The methods used for data analysis; • Inclusion and exclusion criteria; and • An ethical approval statement.

Analysis*	Data Collection and Analysis; Data Analysis (Strategy); Characterization; Statistical Methods/Models/Approach; (Calculations and) Statistics; Empirical Analysis; Regression Results; Analytical Techniques; Data Used; Targeted Statistical Data Analysis; Quantification and Statistical Analysis; Data Analysis and Management; Scaling Analysis; Data Management and Analysis; Analytic Procedure	The analysis section describes how the researcher manipulated the data to obtain their results and can describe both quantitative and qualitative processes. For statistical analysis specifically, this section includes which statistical tests were performed, the sample sizes, the differences among samples, and the kind of statistical software used (name, version, and release number). For qualitative analyses, this section describes how inferences and themes were developed and often references a specific method or paradigm.
Ethical Approval*	Ethics Statement; Statement of Ethics; Ethics/Ethical Approval (Declaration); Ethical Declarations/Clearance; Animal Research Ethics Statement; Animal Welfare Statement; Research Ethics; Statement of Human and Animal Rights; Compliance with Ethical Standards; Ethical Consideration	The ethical approval section is a statement indicating whether or not the researchers have obtained approval from an appropriate institutional review board.
Results	Outcomes; Findings; Research Findings; Implementation Results; Data Analysis; Regression Results; Empirical Analysis	The results section presents a study's findings and observations ideally as neutrally as possible, without bias or interpretation.
Discussion	Discussion of Results; General Discussion; Concluding Discussion	The discussion section puts the results into broader context and establishes their significance in relation to the stated objective, discussing new insights that resulted from the study and explaining any differences or similarities to other published evidence about the topic.
Conclusion*	Summary; Summary and Conclusions; Concluding Remarks	The conclusion is a section that discusses the main takeaways from the study and sometimes implications for changes in research practice or future research opportunities.
Limitations*	Strengths and Limitations; Limitations and Directions for Future Research; Assumptions and Limitations	The limitations section discusses important weaknesses in the design or scope of the study in a way that highlights their consequences for the interpretation of the results. This section places the study in context and addresses the generalizability, applications to practice, and utility of the study's findings. Often, the limitations section discusses opportunities for future research based on the identified weaknesses.
Acknowledgments	Acknowledgment	The acknowledgment section primarily serves to recognize important individuals who made the work possible. This section can also include information about funding, affiliated institutions, associated fellowships, and other miscellaneous information.
Funding Statement*	Funding Sources/Information; Sources of Funding/Support; Financial Disclosure Statement; Financial Support; Research Funding	The funding statement indicates whether or not the authors received funding for their research, describes the role of each funder, and often provides specific grant information and grant numbers.

Table 1. (continued)

Section	Sample alternative labels	Definition
Data Availability Statement*	Data Access(ibility) (Statement); Data Sharing (Plan); Open Practices Statement; Data (Transparency) Statement; Reproducible Research Statement; Data and Materials Availability; Availability of Materials and Data; Availability of (Supporting) Data and Materials; Data Archiving (Statement); Statement on Open Data; Data Policy/Repository/Deposition; Data Submission/Records/Documentation; Open Data Badge; Replication of Results	A data availability statement references a data set that would be necessary to interpret or replicate a study's findings and explains if, how, and under what conditions that data can be accessed.
Code Availability Statement*	(Source) Code; Replication of Results; Open Practices Statement; Reproducible Research Statement; Availability of (Supporting) Source Code and Requirements	A code availability statement references code that would be necessary to interpret or replicate a study's findings and explains if, how, and under what conditions that code can be accessed.
Transparency Statement*	Declaration of Transparency (and Scientific Rigor)	The transparency statement is a standardized declaration in which the lead author affirms that the manuscript is an honest and accurate account of the research, that no important aspects of the study have been omitted, and that any changes to the study were adequately explained.
Open Access Statement*	Open Access (License)	A standardized statement that verifies that the author has followed open access principles in the publication of the study and that specifies the article's open-access license.
Additional Information	N/A	The additional information section is one in which authors disclose competing interests, open access information, funding information and, in some cases, statements indicating the availability of data, code, and software.
Supplementary Information*	Supplementary Material(s); Supplementary Materials and Data; Supplementary Data; Supporting Information	The supplementary information section provides additional information that was not included in the main text but that is important to the scientific integrity of the paper. Traditionally meant to provide information not critical to the main objectives of the research, supplementary information now includes a wide variety of material, including additional figures, tables, methods, background, and citations.
Author Contributions*	Author Information; Statement of Authorship; Authorship Statement; Contributorship; Notes on Contributor(s)	An author contribution statement details each author's role in developing and publishing the manuscript, often following the CRediT format (Brand, Allen et al., 2015), which provides standardized phrases to describe different ways that an author may have assisted with the project.

Conflict(s) of Interest*	Competing Interests (Statement); Competing Interests Declaration; Disclosure (Statement); Declaration of (Competing) Interest(s); Disclosure of Relationships & Activities	A conflict of interest statement acknowledges any financial, legal, commercial, or professional relationships that the researcher or the researcher's employer has with another organization or person that could influence the author's research.
Corresponding Author	Correspondence; Author Contact Information	The corresponding author section provides contact information for one or multiple of the authors.
Publication Details	Article Details	The publication details section lists publication information about the article, including the dates when it was received, accepted, and published, the number of pages, the ISBN or ISSN, and the number of tables and figures.
References	Literature Cited	The references section provides, in a standardized format, publication information about all outside sources of information used to inform the research, giving enough detail that readers can ascertain the genuineness and reliability of the sources.

* This section can be used either as a main section or a subsection.

Table 2. Seven sections, each a subsection of the methods section, were relevant to some but not all articles. Known collectively by 116 labels, these sections could be defined by consistency of use

Section	Alternate labels	Definition
Materials	Materials and Apparatus; Experimental Materials; Apparatus; Instruments; Equipment	The materials section provides a description of any equipment, instruments, software, or other materials used to conduct the research or analysis that would affect the replicability of the work and describes how that equipment was prepared and used.
Study Design	Experimental Design; Design of the Study; Dataset Description; Design; Data and Methods; Research Approach; Research Design; Study Design and Protocol; Procedure; Assumptions; Theoretical Framework; Overview of Experiments	The study design section describes key elements of the study approach, including whether the researcher sought qualitative or quantitative data, what type of measurement framework was employed, and what the units of study were.
Study Subjects	Patients; Participants; Subjects; Study Participants and Recruitment Procedure; Sample; Animals; Population Description and Sample Size; Clinical Data; Participants and Clinical Measures; Study Population; Patient Population; Sampling; Species' Occurrence Data; Study Species; Study Population and Sample Size; Sample Collection	The study subjects section is a description of the people or animals who were studied, including the number of subjects, relevant demographic and health information, and relevant differences among groups of subjects. This section is frequently combined with the "Selection Criteria" section (see below), but they are sometimes presented separately.
Selection Criteria	Clinical Samples; Subject Selection; Dataset Used; Sampling Method; Sampling Techniques; Source Material; Study Design and Sampling; Source of Data; Inclusion Criteria; Patient Selection; Participant Recruitment; Target Samples; Study Design of Patient Analysis; Research Sample; Sample Design; Recruitment and Screening; Selection; Inclusion and Exclusion Criteria	<p>The selection criteria section describes:</p> <ul style="list-style-type: none"> • The eligibility conditions that study subjects or study units had to meet to be included in the study; • Any specific exclusion criteria; and • How the sample size was achieved. <p>In studies in which participants were recruited, this section also describes recruitment and selection methods.</p>

Study Area	Experimental Site; Site Description; Study Site; Geologic Setting; Setting; Study Place; Organizational Setting; Study Location(s); Geological Background; Regional Setting; Study Area and Site Selection; Sites; Area of Study; Study System; Field Site; Case Study Description	The study area section describes where and when the study was conducted, including any information relevant to the specific research question.
Measures	Instrument(s); Experimental Design and Treatments; Variables; Interview Themes; Research Instrument; Questionnaire Design; Study Instrument; Experimental Treatments; Survey Measures; Characterization; Questionnaire Survey; Modeling and Calculations; Computational Method; Survey; Models; Development of the Questionnaires; Measurement; Survey Structure; Tests; Study Definitions; Assessments; Stimuli; Survey Design	The measures section describes the framework and specific methods used to collect and assess data, including a description of the reliability of those methods. In qualitative studies, the measures section often describes a survey or interview instrument. For quantitative studies, the measures section may describe specific equations, models, scales, tests, or surveys used to assess data.
Procedure	Laboratory Analysis; Experimental Layout; Experimental Procedure(s); Data Collection; Preparations; Sampling and Testing; Steps in Data Processing; Method of Data Collection; Data Collection and Sampling; Application of Research; Data; Data Extraction; Tests Performed; Measurement; Method; Tests; Observations; Test Protocols; Data Collection Procedure(s); Process of Data Collection; Sample	The procedure section explains how the measures were applied to collect data and describes any processes to promote data quality.

Table 3. An elaboration upon Tables 1 and 2, consisting of ways in which section and labels may appear in combination

Section 1		Section 2	Labels and alternate versions
Methods	+	Results	Methods and Results
Methods	+	Statistical Analysis	Data Analysis, Enquiry, Methodology, & Applications
Methods	+	Materials	Materials and Methods; Patients, Materials, and Methods; Approach; Experiment(al)
Methods	+	Discussion	Implementation and Discussion
Statistical Analysis	+	Results	Analysis and Results
Discussion	+	Statistical Analysis	Analysis and Discussion
Discussion	+	Results	Results and Discussion; Findings and Discussion
Discussion	+	Conclusions	Discussion; Conclusions and Discussion; Conclusions, recommendations and suggestions
Funding	+	Acknowledgments	Acknowledgments and Funding
Code Availability Statement	+	Data Availability Statement	Code and Data Availability Statement; Data and Code(s) Availability Statement; Availability of data and materials; Availability of materials and data; Replication of Results; Open Practices Statement; Data and Software Availability

combined section includes the rhetorical components of each of the individual sections but is listed under only one label.

Another notable phenomenon that was encountered was the tendency to use individualized labels, or labels that were specific to the article’s topic. These individualized labels were particularly common for the Background section. For instance, one article (Vlachantoni, 2019) labeled its Background, “Conceptualising need for social care,” and another (Brooks, Tejedo, & O’Neill, 2019) labeled its Background, “General characteristics of Antarctic soils.”

Although the project team encountered significant variation among RAs, particularly in the specific labels used to demarcate sections, the project team was able to use the patterns in the data to create a model of a sample RA, shown in Figure 4. This model represents the typical order and organization of the various RA sections. Importantly, some articles strayed notably from the model. In particular, RAs that focused on a proof of concept tended to have few labeled sections, instead presenting models, equations, and results in a thematic order. Physics and mathematics articles tended to follow this thematic pattern.

3.2. Objective 3: Identifying Research Articles

Our third objective was to determine whether the sections and labels can be used to consistently determine whether an article *is* an RA or whether an article is *not* an RA. Achieving this objective required determining the frequency of use of each of the sections in RAs and comparing those results to the frequencies in nonresearch articles. It also required the identification of any sections that were unique to either RAs or nonresearch articles.

The research team found that certain sections were nearly universal across RAs, such as Abstract (99.6%), Introduction (89.2%), Methods (97.2%), Results (98%), and Discussion (92%). The only truly universal section across all RAs was the References section (100%).

Table 4. Frequency with which each section appeared in traditional RAs and nontraditional articles, including a *p*-value when the researchers found a statistically significant difference

	Research article total count (<i>n</i> = 251)	Nonresearch article total count (<i>n</i> = 30)	Research article percentage	Nonresearch article percentage	<i>p</i> -value
Abstract	249	30	99.6	100.0	
Objective	19	0	7.6	0.0	
Introduction	223	28	89.2	93.3	
Background	32	4	12.8	13.3	
Statistical Analysis	110	9	44.0	30.0	
Analysis	139	15	55.6	50.0	
Materials	21	2	8.4	6.7	
Measures	45	6	18.0	20.0	
Study Design	42	1	16.8	3.3	
Procedure	76	13	30.4	43.3	
Study Location	45	1	18.0	3.3	
Ethics Statement	61	8	24.4	26.7	
Acknowledgments	173	18	69.2	60.0	
Limitations	46	3	18.4	10.0	
Transparency Statement	7	1	2.8	3.3	
Funding Statement	89	10	35.6	33.3	
Corresponding Author	107	8	42.8	26.7	
Author Contributions	63	8	25.2	26.7	
Publication Details	24	2	9.6	6.7	
Conflict of Interest	155	14	62.0	46.7	
Data Availability Statement	41	3	16.4	10.0	
Code Availability Statement	0	0	0.0	0.0	
Methods	243	21	97.2	70.0	6.07×10^{-8}
Study Subjects	68	2	27.2	6.7	2.63×10^{-2}
Selection Criteria	24	10	16.8	33.3	5.07×10^{-4}
Results	245	16	98.0	53.3	1.36×10^{-17}
Discussion	230	19	92.0	63.3	1.65×10^{-5}
Results and Discussion	38	0	15.2	0.0	0.0445
Conclusion	171	14	68.4	46.7	0.0324
References	250	27	100.0	90.0	7.23×10^{-4}
Open Access Statement	73	2	29.2	6.7	0.0161

Other sections were unique to a particular publisher or were infrequently used, such as Publication Details (9.16%), which was only encountered with one publisher, and Background (12.8%). For a full list of results on section frequency, see Table 4.

Although some significant differences between RAs and meta-analyses, case studies, and review articles were found in each major section in RAs, they could also be found in nonresearch articles. As a result, the presence of one of these sections cannot be easily used to distinguish RAs from other journal article types.

There were specific sections that were found only in nonresearch articles, specifically “Case Study,” and “Meta-analysis.” Though these sections were not in every case study or meta-analysis, every paper that included either of these could quickly be identified as either a case study or meta-analysis. Some nonresearch articles use different labels to refer to a given section. For instance, many review articles referred to the Selection Criteria section as “Search Strategy,” “Search Methods,” or “Study Selection.” These specific terms were not used in RAs and could therefore be used to determine if a journal article is *not* an RA.

4. DISCUSSION

The objectives of this study were to create a normalized set of journal sections and labels, determine structural differences between RAs and nonresearch articles, and determine whether sections can be used to identify an RA as traditional among a constrained corpus of articles within a set of specific publishers and journals. The results of our inquiry have significant implications regarding the initial issues that we set out to solve: difficulty identifying journal articles as RAs and difficulty querying RAs to locate particular types of information across a corpus of journals and publishers.

The different forms for RAs between publishers and among journals add a further dimension to the work began by Thelwall (2019). Not only did these results uphold this previous work, but further showed the differentiation among journals and publishers. It may be expected that a specific publisher would apply similar standards and requirements for how RAs are formatted, yet this was not found to be true. Taking this a step further, within a specific journal you may expect to see consistency with how RAs are structured, yet this was also found to be inconsistent.

With regard to identifying RAs, it was found that RAs and similar genres, such as meta-analyses, review articles, and case studies, cannot be easily distinguished based on the major RA sections alone: A, I, B, M, R, D, or C. Furthermore, these article types tend to include similar rhetorical components or moves. For instance, Kanoksilapatham (2015) identifies the Results moves of an RA as “summarizing procedures,” “reporting results,” and “commenting on results,” moves that are common to all article types. Instead, differentiation may occur at the step level. For instance, one Introduction move is to “[establish] a territory to provide background information of the research topic,” which may be present in all article types, but a review article may not include the typical step of “claiming centrality.” There is thus more promise in differentiating journal article types by their subsections rather than by their main sections.

Another potential way to differentiate journal article types is by the specific labels used to identify a given section. For instance, both review articles and RAs frequently include Selection Criteria sections, which is defined as a section that “describes: 1) The eligibility conditions that study subjects or study units had to meet in order to be included in the study; 2) Any specific exclusion criteria; and 3) How the sample size was achieved.” In studies in which participants were recruited, this section also describes recruitment and selection methods. Currently, there are far too many

variations on these labels for them to be a practical way to identify RAs or non-RAs, but if these labels were standardized by genre it would be much simpler to use machines to accurately sort journal articles by genre or to find an algorithm that can identify a combination of these.

In terms of querying articles, the extensive variation in section labels is a significant barrier to comprehension for both human and nonhuman readers. In particular, the subsections of the Methods section could be very convoluted, especially because different authors used the same labels to refer to different types of information. For example, “Research Approach” was used to refer to a study’s data collection methods in one article but “Study Design” in a different article. Similarly, “Conflict of Interest” and “Ethics Statement” were often used interchangeably, but “Ethics Statement” also often referred to approval from an Institutional Review Board. This type of inconsistency makes comprehension a challenge for human readers as well as machine-based readers. And, for machine readers in particular, the sheer number of possible labels and authors’ tendency to use individualized labels makes it almost impossible to identify sections based on their labels. This reality is a problem particularly for researchers who wish to conduct section-based analyses.

It is important to note again the tension between clarity and accuracy. By nature, RAs present new and often complicated information, and standardized section labels could be an important way for authors to signal what type of information they will be presenting. Moreover, much of the variation did not serve any significant rhetorical purpose and would not, therefore, decrease accuracy. For instance, the difference between “Materials and Methods” and “Methods and Apparatus” is trivial. Unnecessary variations such as these, which can significantly impede machine-based analysis, should be minimized through a normalized set of labels and definitions agreed upon by the scientific community. Such an intervention would not only facilitate machine-based analysis but would also ease other researchers’ ability to replicate and understand a study’s findings and processes. Researchers could begin the process of standardization by reviewing the author guidelines provided through the [Equator Network \(2006\)](#), which outline best practices for RA sections but fall short of suggesting label names or precise definitions.

5. LIMITATIONS

There are a few notable limitations of our study. First, although we looked across the journals of 10 major publishers, an expanded list of publishers could have significantly altered our results. In our analyses, we found distinct patterns within both journals and publishers, so a different set of publications could have yielded very different results. Another limitation of this study is that we only harvested open access articles. A different set of researchers may be drawn to open access publications and could affect how those articles are structured. Perhaps the most significant limitation of this study was that we had a limited sample of meta-analyses, review articles, and case studies and that we grouped these types of articles together despite their differences. A more nuanced analysis could compare the sections in RAs to just one of these other article types to see if there are more consistent ways to differentiate RAs from specific subtypes. In this article, we grouped these subtypes together because we hoped to find an easy way to differentiate RAs from all other journal article types, but we did not find a way to do so. Individual analyses may therefore have been more appropriate. Finally, there was unavoidable subjectivity built into our study. Particularly in the methods section, where subsections often had convoluted purposes, the research team had to make decisions about the overall rhetorical purpose of that section and make the appropriate classification. We also had to decide which labels to choose as the official labels for various

sections, decisions which could reasonably be debated. Our work should therefore be viewed as a starting point for future research, not the final answer to our questions.

6. CONCLUSION

In this study, 31 different RA sections known by 302 different labels were identified across publishers and journals. Establishing agreed-upon names and definitions for these sections is an important aspect in any attempt to improve human and nonhuman readers' ability to identify, interpret, or analyze the results of RAs. Although nuance and individualized section labels can sometimes provide greater accuracy, the scientific community must consider the ways in which specificity can undermine comprehension by machines. These questions are particularly important because the RA is a genre concerned with communicating new and complicated ideas. Furthermore, RAs are based on the scientific method, a highly structured process. Greater standardization in RAs, with labels that describe particular parts of the scientific method, could greatly improve comprehension and analysis. Of course, flexibility must remain, and standardization must always be balanced with accuracy. Nonetheless, researchers must reckon with the reality that their research could be overlooked, especially in machine analyses, if they stray too far from standard practice.

This study can be an impetus for journals and publishers to adopt agreed upon section labels and definitions and provide more guidance for authors about article structure. No matter the specific outcomes, the scientific community must begin discussing how the current form of the RA affects comprehension, reproducibility, and analysis, especially in a new age of machine-based analysis.

AUTHOR CONTRIBUTIONS

Sarah Nathan: Conceptualization, Data curation, Formal analysis, Writing—original draft, Writing—review & editing. Leah Haynes: Conceptualization, Data curation, Formal analysis, Writing—original draft, Writing—review & editing. Jessica Meyer: Conceptualization, Data curation, Formal analysis, Writing—original draft, Writing—review & editing. Josh Sumner: Conceptualization, Data curation, Formal analysis, Writing—original draft, Writing—review & editing. Cynthia Hudson Vitale: Conceptualization, Study supervision, Writing—original draft, Writing—review & editing. Leslie D. McIntosh: Conceptualization, Study supervision, Writing—original draft, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

No funding has been received for this research.

DATA AVAILABILITY

Data and statistical analysis are publicly accessible on Figshare and citable as: Vitale, C., Nathan, S., Haynes, L., Meyer, J., Sumner, J., & McIntosh, L. D. (2021). Dataset for manuscript: An analysis of form and function of a research article between and within publishers and journals. Figshare. <https://doi.org/10.6084/m9.figshare.14502168>.

REFERENCES

- American Association for the Advancement of Science. (n.d.). *Instructions for preparing an initial manuscript*. <https://www.sciencemag.org/authors/instructions-preparing-initial-manuscript>
- Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing, 28*, 151–155. <https://doi.org/10.1087/20150211>
- Brooks, S., Tejedo, P., & O'Neill, T. (2019). Insights on the environmental impacts associated with visible disturbance of ice-free ground in Antarctica. *Antarctic Science, 31*(6), 304–314. <https://doi.org/10.1017/S0954102019000440>
- Dahl, T. (2009). The linguistic representation of rhetorical function: A study of how economists present their knowledge claims. *Written Communication, 26*(4), 370–391. <https://doi.org/10.1177/0741088309341241>
- de Waard, A., & Pander Maat, H. (2012). Verb form indicates discourse segment type in biological research papers: Experimental evidence. *Journal of English for Academic Purposes, 11*(4), 357–366. <https://doi.org/10.1016/j.jeap.2012.06.002>
- Dobakhti, L. (2013). Commenting on findings in qualitative and quantitative research articles' discussion sections in applied linguistics. *International Journal of Applied Linguistics and English Literature, 2*(5). <https://doi.org/10.7575/Aiac.ljalel.v.2n.5p.145>
- Equator Network. (2006). <https://www.equator-network.org/>
- Hsieh, W. M., Tsai, T. C., Lin, M. C., Liou, H. C. U., & Kuo, C. H. (2006). Exploring genre sets: Research article sections in illustrative humanities and science disciplines. In *The 23rd International Conference on English Teaching and Learning in the ROC* (pp. 249–262). Taipei.
- ICMJE. (2004). Uniform requirements for manuscripts submitted to biomedical journals: Writing and editing for biomedical publication. *Haematologica, 89*(3), 264.
- Journal of Environmental Quality. (2021). Author instructions: JEQ manuscript submission template. <https://www.agronomy.org/publications/journals/author-resources/jeq-instructions/>
- Kanoksilapatham, B. (2015). Distinguishing textual features characterizing structural variation in research articles across three engineering sub-discipline corpora. *English for Specific Purposes, 37*, 74–86. <https://doi.org/10.1016/j.esp.2014.06.008>
- Kashiha, H. (2015). Recurrent formulas and moves in writing research article conclusions among native and nonnative writers. *Southeast Asian Journal of English Language Studies, 21*(1). <https://doi.org/10.17576/3L-2015-2101-05>
- Li, L.-J., & Ge, G.-C. (2009). Genre analysis: Structural and linguistic evolution of the English-medium medical research article (1985–2004). *English for Specific Purposes, 28*(2), 93–104. <https://doi.org/10.1016/j.esp.2008.12.004>
- Lin, L., & Evans, S. (2012). Structural patterns in empirical research articles: A cross-disciplinary study. *English for Specific Purposes, 37*(3), 150–160. <https://doi.org/10.1016/j.esp.2011.10.002>
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*, 2nd Edn. London: Chapman & Hall. <https://doi.org/10.1007/978-1-4899-3242-6>
- Nwogu, K. N. (1997). The medical research paper: Structure and functions. *English for Specific Purposes, 16*(2), 119–138. [https://doi.org/10.1016/S0889-4906\(97\)85388-4](https://doi.org/10.1016/S0889-4906(97)85388-4)
- PLOS ONE. (n.d.). *PLOS ONE submission guidelines*. <https://journals.plos.org/plosone/s/submission-guidelines#loc-results-discussion-conclusions>
- Prodigy. (2017). Retrieved from <https://prodi.gy/>
- Pruim, R., Kaplan, D. T., & Horton, N. J. (2017). The mosaic package: Helping students to 'think with data' using R. *The R Journal, 9*(1), 77–102. <https://journal.r-project.org/archive/2017/RJ-2017-024/index.html>. <https://doi.org/10.32614/RJ-2017-024>
- R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. <https://www.R-project.org/>
- SpaCy. (2017). Retrieved from <https://spacy.io/>
- Swales, J. (1981). *Aspects of article introductions*. Birmingham, UK: University of Aston.
- Tanti, M. (2014). Towards a definition of the scientific paper? *4th International Symposium ISKO-Maghreb: Concepts and Tools for knowledge Management (ISKO-Maghreb)*, Algiers, (pp. 1–8). <https://doi.org/10.1109/ISKO-Maghreb.2014.7033485>
- Teufel, S. (1999). Argumentative zoning: Information extraction from scientific articles. University of Edinburgh. <https://era.ed.ac.uk/handle/1842/11456>
- Tessuto, G. (2015). Generic structure and rhetorical moves in English-language empirical law research articles: Sites of interdisciplinary and interdiscursive cross-over. *English for Specific Purposes, 37*, 13–26. <https://doi.org/10.1016/j.esp.2014.06.002>
- Thelwall, M. (2019). The rhetorical structure of science? A multidisciplinary analysis of article headings. *Journal of Informetrics, 13*(2), 555–563. <https://doi.org/10.1016/j.joi.2019.03.002>
- Vlachantoni, A. (2019). Unmet need for social care among older people. *Ageing and Society, 39*(4), 657–684. <https://doi.org/10.1017/S0144686X17001118>