8-8-2023

Writing Science for Diverse Audiences: A Corpus-based Discourse Analysis of the Language of Science News and Research Articles

Jordan Batchelor

Writing Science for Diverse Audiences: A Corpus-based Discourse Analysis of the Language of Science News and Research Articles

by

Jordan Batchelor

Under the Direction of Viviana Cortes, PhD

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2023

ABSTRACT

Despite the historic prevalence of the research article (RA) genre in the English for Academic Purposes (ESP) literature, work examining the ways that academic research is communicated with broader audiences—sometimes referred to as ‘popularization’ or ‘popular science’ (Gotti, 2014)—is on the rise. Scholars from diverse fields have shown interest in contexts of popular science in part because they represent a meeting point between the general public and academia. However, much of the research examining the language of popular science has adopted a rhetorical rather than linguistic lens (Pérez-Llantada, 2021). In addition, the recent interest in digital multi-modal genres (e.g., Luzón, 2023; Xia, 2023) has left the linguistic features of written discourse comparatively under-examined, and studies adopting corpus approaches have often included texts which are out-of-date, few in number, or under-described with regard to their place under the umbrella of popular science.

This dissertation applies a mixed methods design to a new corpus representative of one variety of popular science writing, namely online science news articles (SNAs). It uses computer programs to compare the linguistic profiles of 400 SNAs with a matching corpus of the 400 RAs. Specifically, this dissertation investigates features of the verb phrase, namely short verb phrase variation, long verb phrase variation, and attribution of knowledge via reporting clauses. These features offer a useful contrast to the current noun-focused approach to grammatical complexity research (see Lan, Liu, & Staples, 2019). To inform interpretation of corpus findings, discourse-based interviews (Conrad, 2014) with seven SNA writers were also employed.

Findings from the linguistic analyses, analyses of the registers’ situational characteristics, and informant interviews highlight the many differences between the registers, differences motivated especially by characteristics of audience, textual layout, and purpose. SNAs are short texts which function to inform and entertain an audience of mixed expertise. As a result, they utilize more verbs overall, as well as features of short and long verb phrases which allow writers to report research activities as stories involving researchers, their beliefs, and their words. Implications relating to contexts of science communication and pedagogical applications are discussed.

INDEX WORDS: Popular science, Research articles, Science communication, The verb phrase, Corpus linguistics, Science journalism

Writing Science for Diverse Audiences: A Corpus-based Discourse Analysis of the Language of Science News and Research Articles

by

Jordan Batchelor

Committee Chair: Viviana Cortes

Committee: Diane Belcher

Eric Friginal

Carmen Pérez-Llantada

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

August 2023

DEDICATION

This dissertation is dedicated to my parents, David and Doreen.

ACKNOWLEDGEMENTS

There are many people who deserve to be acknowledged for helping me reach this goal. First, I will always be grateful to my parents, who have supported whatever I have wanted to do, even when I wanted to get a PhD in philosophy. There are also colleagues and mentors at Pennsylvania State University, which I attended for my Master's degree, who encouraged and supported my desire to apply for PhD programs. Penn State is also the place where I met my wife, Meng. Meng has endured the many good and bad times with me over the last four years, and I hope to pay her back in the future. I would also like to thank the professors and fellow students in the Department of Applied Linguistics & ESL at Georgia State University. In particular, my advisor and mentor, Dr. Viviana Cortes, has spent countless hours helping me prepare for comprehensive exams, develop this dissertation, and generally grow into the applied linguist that I am today. Similarly, I am grateful to my dissertation committee members, who took time out of their busy professional lives to participate in the proposal and defense of this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviations for 14 verb patterns

IV+0	Intransitive pattern
FCl+S|TV:	Reporting clause patterns
Ex*There*+*BE*:	Existential *there* pattern
Question	Interrogative pattern
CV+PhC	Copular patterns with phrasal complements
CV+FCl	Copular patterns with finite complement clauses
CV+NfCl	Copular patterns with non-finite complement clauses
PassTV	Passive patterns without a complement clause
PassTV+FCl	Passive patterns with a finite complement clause
PassTV+NfCl	Passive patterns with a non-finite complement clause
TV+PhO	Transitive patterns with phrasal objects
TV+FCl	Transitive patterns with finite complement clauses
TV+NfCl	Transitive patterns with non-finite complement clauses
TV+PhO+NfCl	Transitive patterns with phrasal objects and non-finite complement clauses

1 INTRODUCTION

Research genres hold a critical place in the field of English for Specific Purposes (ESP), particularly its sub-field of English for Academic Purposes (EAP). ESP refers to the teaching and learning of English in specific domains (Paltridge & Starfield, 2013), while EAP is one domain which has garnered particular attention (Anthony, 2018), due in part to the central role of English in teaching, learning, and sharing academic knowledge (Lillis & Curry, 2010). While EAP researchers have examined a wide variety of contexts, one of the most well represented is post-secondary learners and their communicative needs (Belcher, 2013), resulting in a significant body of literature examining the English used in higher education and related contexts. It is generally recognized that teaching to the specific needs of learners leads to better pedagogical outcomes (Hyland, 2002a), opening the door to research on the language of academic lectures, argumentative essays, textbooks, PhD defenses, conference posters, and more, for the purpose of improved pedagogy and improved understanding of the language used in those contexts.

No academic genre has garnered more attention than the research article (RA) (Samraj, 2016), and for good reason. Academia in recent decades has produced—and is continuing to produce—more published research articles in more academic journals across more subject area specializations than ever before (Hyland, 2022). Thus, the relevance of the RA genre is only increasing. The rhetorical and linguistic complexity of research writing means that producing a successful RA is not self-evident, either for those writing in their primary language or those writing in an additional one. Swales' (1981, 1990) influential genre analysis has motivated a vast body of literature exploring the 'moves' and 'steps' of RAs and its part genres. Additionally, the recognition that research writing is often influenced by the discipline in which one writes

(Hyland, 2007) has essentially multiplied the possible research questions that researchers can ask and answer within the framework of EAP.

At the same time, the RA represents only one genre among many that academics produce and audiences consume. In recent years, language scholars have shown increasing interest in scientific genres wherein the characteristics of author and audience are more broadly inclusive. Whereas RAs are typically written by academics *for* academics, often within a specialized field, communicative events like science magazines, science blogs, and science videos are produced with broader audiences in mind, resulting in different rhetorical and linguistic strategies to meet the needs of their audience. Moreover, thanks to advances in computer technologies, these kinds of communicative events are expanding (Pérez-Llantada, 2016).

The recontextualization of scientific information for broad audiences is sometimes referred to as *popular science* (Calsimiglia & Van Dijk, 2004; Gotti, 2014), a concept which has been primarily examined by scholars working within the history and sociology of science, media studies, and linguistics. Historians have sought to understand the social conditions under which popular science has developed, turning in particular to Victorian-era Britain (Lightman, 2007). Initially, working scientists communicated the scientific theories of their day to the public themselves, but this increasingly became the task of journalists as science split into increasingly esoteric specializations (Dunwoody, 2014). Thus, publications such as *Scientific American* and *Popular Science* (both U.S.-based periodicals) became avenues through which non-scientists could be informed about and entertained by the world of science.

Popular science writing in mass media continued throughout the 20th century, with various ebbs and flows of public interest. However, scholars began to critique the popular conception of popular science, which states that popularization vulgarizes and simplifies genuine science for consumers unable to directly consume real science (Hilgartner, 1990; Whitley, 1985). In particular, the role and characteristics of the audience has been contested, with scholars noting difficulties in clearly distinguishing between scientific and general audiences (Myers, 2003). Nonetheless, contemporary research of popular science often notes the importance of audience characteristics when examining science communication, a feature which has only increased in relevance in the digital era (Trench, 2008).

More scholars are exploring science discourse communicated online and the genres through which it is communicated, exampled by the number of recent studies, theoretical articles, and edited volumes on the topic (e.g., Belcher, 2023; Liu et al., 2023; Luzón & Pérez-Llantada, 2019). Most relevant to this dissertation, the area of science communication research which examines the rhetorical and linguistic features of texts, often explored within the framework of EAP, is also increasing. For example, scholars have examined the ways that TED Talks and Three-Minute Thesis presentations may offer rich resources for English listening practice or developing academic presentation skills (e.g., Liu, 2023). Similarly, popular science writing offers its own strengths as learning material through its use of lexical, grammatical, and rhetorical features that accommodate non-technical readers. Thus, the exploration of digital genres communicating academic information to broader audiences rightfully fits within the paradigm of EAP research.

However, there remain gaps in this literature to which new research can contribute. The increased attention in digital genres communicating science have rightfully paid close attention to new genres native to the Internet and the affordances that they bring. In some cases, the technological innovation of the genre, as opposed to its popularization, is the focus, such as with digital or graphical research abstracts. In other cases, attention has been paid toward genres in which the academic is also the text's author, as is the case of research blogs and science-focused crowdfunding proposals. These communicative contexts are deserving of attention, but it is noteworthy that traditional written genres of popular science have garnered less attention in recent years. In particular, few recent studies have examined journalistic genres of written science communication in the digital age. Of those that have, several utilized publicly available corpora consisting of 20th century texts, while others compiled private corpora ranging significantly in purpose and generalizability. However, the digital age has increased the presence of traditional written genres of popular science, such as news articles authored by journalists in online periodicals. This dissertation, thus, seeks to fill this gap by examining the language of contemporary science news writing online authored by science writers rather than the scientists themselves. Additionally, since this flavor of popular science reports on recently published professional science, namely research articles, it offers a unique comparison between the language of popular and professional science writing. Thus, this dissertation not only offers insight into the language of a corpus of contemporary popular science writing, but it also offers direct comparisons between the popular articles and the professional articles that the popular articles reported on.

As recent studies of science communication online have emphasized the rhetorical features of these genres (Pérez-Llantada, 2019, p. 76), this dissertation takes a comparatively formal linguistic approach, examining grammatical and lexico-grammatical features of written texts. Section 3.1.3 of Chapter 3 describes these features in more detail. In short, this dissertation focuses on three sets of features that revolve around verbs, including patterns of the long verb phrase, variation of the short verb phrase, and reporting clauses and their reporting verbs. These features provide insight into a critical feature of English grammar that has rich functional connections in discourse. Finally, in addition to employing textual methods to examine the language of written texts, discourse-based interviews (Conrad, 2014; Hyland, 2007) with writers of these popular science articles are also adopted to inform interpretations of the textual findings.

The research questions that this dissertation seeks to answer include the following:

1. What are the situational characteristics that differentiate one register of popular science writing, namely online science news articles, from one register of professional science writing, namely published research articles?
2. How does the use of verb patterns vary across the registers?
    a) How does the use of broad and specific valency patterns vary across the registers in terms of frequencies and functions?
    b) How do writers of online science news articles explain the use of these features in their texts?
3. How does variation in the short verb phrase compare across the registers?
    a) How does the use of tense, aspect, modality, and voice vary across the registers in terms of frequencies and functions?
    b) How do writers of online science news articles explain the use of these features in their texts?
4. How does the use of reporting clauses vary across the registers?
    a) How does the use of reporting clauses to attribute knowledge vary across the registers?
    b) How does the choice of reporting verbs vary across the registers?
    c) How do writers of online science news articles explain the use of these features in their texts?

In adopting these questions, I intend to contextualize this dissertation within the framework of EAP, providing linguistic insights into texts which may serve as useful classroom materials, though while also having implications for non-pedagogical perspectives. For example, an examination of contemporary science news writing may also be of interest to scholars working in science communication (e.g., Bucchi & Trench, 2014), genre theory (e.g., Luzón & Pérez-Llantada, 2019) and science journalism (e.g., Bauer & Bucchi, 2007).

This dissertation is organized as follows. In Chapter 2, I review relevant literature on popular science, including a brief description of relevant features of its history (2.1), a review of linguistic and rhetorical studies of science communication (2.2), and a description of the gaps that this dissertation intends to fill (2.3). Chapter 3 describes the methodology, including a description of the corpus, corpus tools and procedures, and feature selection and analysis (3.1), as well as a description of the discourse-based interviews (3.2). Chapters 4, 5, 6, and 7 report on the findings organized by research question. Chapter 4 reports on the situational characteristics analysis, followed by chapters pertaining to the analysis of verb patterns (Chapter 5), short verb phrase variation (Chapter 6), and reporting clauses (Chapter 7). Finally, Chapter 8 concludes the dissertation by offering a review of the findings, their implications, and the limitations of the dissertation. References and appendices can be found after the conclusion chapter.

2 LITERATURE REVIEW

Contemporary research of popular science often highlights the recent proliferation of its discourse and the genres through which it is communicated, owing especially to advancements in computer technology (Pérez-Llantada, 2016). However, neither the notion of popular science nor its study was the result of computers. Rather, scientific work has been popularized for centuries and the academic study of it reaches back decades before contemporary technologies. Throughout its study, media scholars, historians, and sociologists of science have contributed substantially to the history, theory, and description of popular science. Thus, I begin this literature review by providing a brief review of this literature in section 2.1 to set the context for the linguistic study of popular science, which is covered more substantially in section 2.2. Finally, I conclude the literature review by highlighting the gaps in the literature which this dissertation intends to fill in section 2.3.

2.1 A brief history of popular science

Historians identify one of the earliest examples of science popularization as Bernard Le Bovier de Fontenelle's *Conversations on the Plurality of Worlds*, published in France in 1686. This brief fictional work was written in French rather than Latin, attempted to explain current scientific theory for a non-technical audience, and mused about the possibility of life beyond Earth. Prior to this time, popularizing such views would have been controversial and potentially dangerous, since it challenges the then theological beliefs about humanity's place in the universe (Fontenelle, 1990, p. vii). In the years following de Fontenelle, shifting ideologies allowed for a greater presence of scientific discourse, bringing such topics as Newtonian physics, chemistry, and electricity into public spaces (Bensaude-Vincent, 2001, p. 102).

By the 1800s, science underwent a process of professionalization, which included the development of salaried positions, formal training opportunities, and specialized publications for research (Barton, 2003, p. 76). Importantly, the professionalization of science also led to more *popularization* of science. As scientists specialized in increasingly esoteric fields, there arose a greater need for communicators who could translate the professional world for the public one (Lightman, 2007, p. 495). Thus, just as one could forge a career as a scientist, so too could they inform and entertain the public as a popularizer. Entering the 20th century, revolutions in physics and biology, as well as the growing presence of electricity in modern life, accelerated this gap between scientists and the public, and popularization increasingly became the job of journalists rather than the scientists themselves (Dunwoody, 2014).

A series of events and innovations beginning with the conclusion of World War II sparked a considerable demand for science understanding in the 20th century, leading to more substantial coverage in print media, greater interest among readers, and the establishment of science journalism as a sub-field within journalism (Bauer et al., 1995; Dunwoody, 2014). Initially, science journalism focused on the physical sciences but shifted toward health and medicine as the century wore on (Bauer, 1998). Although journalists provided a hungry public with greater access to the world of science, it was not always positively evaluated by scientists themselves. Some criticized science news for being inaccurate in its reporting, and others noted that it could be prone to omission of important information, such as research methodology and qualifiers about the generalizability of findings (Pellechia, 1997), perhaps the result of news tending to be either positive/celebratory or negative/critical but rarely mixed (Bauer et al., 1995; Einsiedel, 1992).

At some point during this time, a stereotype that popular science simplifies and distorts actual science developed, a view which sociologists and other researchers sharply critiqued. One of the earliest works to describe this view was Whitley (1985), who referred to the stereotype as the ‘traditional view,’ similar to Hilgartner’s (1990) dominant view and Grundmann and Cavaillé’s (2000) canonical model. These views see the audience of popular science as large, homogenous, and incapable of directly understanding the esoteric knowledge of scientific communities. In doing so, the traditional view provides professional scientists the privileged position of holding genuine knowledge, as well as the ability to delegitimize popular knowledge.

Hilgartner, however, argues that it is rarely possible to make a clear and reliable distinction between genuine and popular science, noting that scientific knowledge can be better described as a cline from ‘upstream’ to ‘downstream’ science, which is illustrated in Figure 2.1.

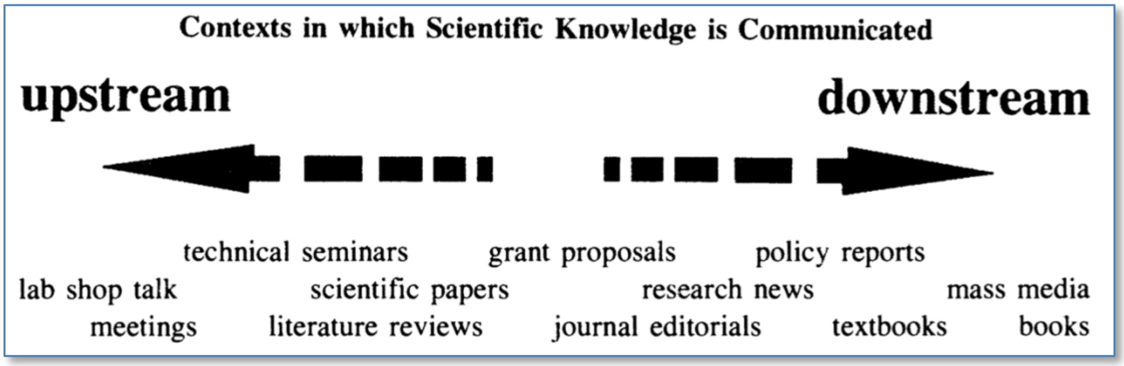


Figure 2.1 The stream of scientific communication from professional to popular, reproduced from Hilgartner (1990, p. 528)

On the one hand, the contexts shown in Figure 2.1 can be roughly divided into genuine and popular science by describing certain characteristics of the contexts. For example, professional science may include those events where the author and audience are knowledgeable specialists, while the remaining events are instances of popular science. However, this enterprise quickly breaks down when applying the distinction across all contexts, especially those closer to one another on the cline (Myers, 2003).

Nonetheless, scholars often rely on such characteristics to differentiate popular from professional science, specifically that the audience tends to be large, general, and non-expert (e.g., Bucchi & Trench, 2014; Calsamiglia, 2003; Fahnestock, 1986; Gotti, 2014; Myers, 1989). The inclusivity of this feature makes the identification of popular science simpler, as any context which communicates scientific information to more than just other scientists can be labeled as popular science. For example, the technology column in a local newspaper, the non-fiction aisle of a bookstore, a social media post, and a science podcast may all communicate scientific information to diverse audiences and thus be examined as instances of science popularization.

In addition to audience, it is also assumed that the language of popular science must diverge at various levels from that of professional science. At the most general level, popular science must recontextualize the technical nature of scientific research for lay readers and listeners (e.g., Calsamiglia & Van Dijk, 2004; Gotti, 2014; Hilgarnter, 1990), resulting in different features and strategies used to communicate its content. This general observation has motivated a body of literature from discourse analysts, applied linguists, and other scholars to better understand how the process of popularization affects linguistic choices. It is this body of literature that I turn to next, in section 2.2, to outline what these studies have revealed about the language of popular science discourse.

2.2 The language of popular science discourse

Owing to advancements in computer technology, the landscape of science communication in the 21st century is very different from that of the previous two centuries. The exponential increase in the amount of published, scholarly work in past decades (Hyland, 2022) has meant that there is now more science to popularize than ever. In turn, many traditional avenues for popularizing science, such as newspapers and magazines, have migrated to the Internet, and new means through which to communicate science are constantly expanding (Luzón & Pérez-Llantada, 2019). Perhaps as a result, more scholars have become interested in exploring the language of science communication, both in their traditional forms and, especially, their new digital forms.

Below, I review this body of literature in three parts. First, I briefly describe important concepts in the style of (professional) 'scientific' writing, providing a context with which popular science writing can be compared and contrasted (section 2.2.1). Second, in section 2.2.2, I review recent work on popular science communicated through digital genres native to the Internet, such as academic blogs, 3-Minute Thesis presentations, and TED Talks. Finally, in section 2.2.3, I review work on popular science writing published in traditional print genres and their digital forms, such as newspapers, magazines, and non-fiction books. Section 2.4 then expands on the gaps in this literature that this dissertation intends to fill.

2.2.1 The discourse style of 'scientific' writing

Before examining literature on popular science discourse, it is worth reviewing some key concepts about 'scientific' discourse. The style of scientific discourse, which here largely refers to written discourse, is important because it is said to contrast with that of popular science (Myers, 1999, p. 141). For example, Adams-Smith (1987) noted the way that some popular medical articles altered their organization, syntax, and vocabulary to suit their general audience, including an increased focus on findings, less complex noun phrases, and more general vocabulary. These features in turn highlight some of the traditional characteristics of scientific writing, typified by impersonal, detailed descriptions of experiments using technical or abstract vocabulary. Indeed, Halliday and Martin's (1996) book on scientific English begins by noting the way that children easily recognize the intimidating style of scientific communication, often being put off by it.

For Halliday and Martin (1996), one of the most characteristic features of scientific writing is its tendency to turn non-nouns into nouns (a type of 'grammatical metaphor'). Thus, 'processes' (verbs) and 'qualities' (adjectives) are expressed as 'things' (nouns), a style of discourse discordant with the way children learn English and people use it in everyday settings. Indeed, others have noted the nominal style of academic writing (Wells, 1960), as well as the tendency to create lengthy, complex noun phrases by way of stacked modifiers (Biber et al., 1999, Ch. 8). This style developed over hundreds of years to suit the needs of communicating scientific experimentation, especially to technical audiences (Halliday & Martin, 1996, p. 74).

These features are not just stylistic but also functional. Biber and Gray (2016) argue that this style allows for compact, informationally dense prose useful when writing for increasingly specialized audiences (p. 207). Indeed, editors of popular science publications are likely aware of the discord between scientific and non-scientific writing. Myer's (1999) investigation of writing popularizations revealed that some editors re-package dense noun phrases and nominalizations as easier-to-read finite clauses (p. 179-180), reversing the grammatical process developed for communicating experimental science. From a sociological perspective, scientific writing values the 'empiricist repertoire,' which denies the writer's subjective interpretation while portraying science as the neutral and inevitable outcome of standard procedures (Gilbert and Mulkay, 1984). Moreover, its opacity arguably performs a gatekeeping role for the world of science, serving to manage who can take part in its discussion (Bazerman, 1988, p. 294). Regardless of the functional explanation, it is clear that the discourse of scientific writing is not random, haphazard, or merely personal preference.

In contemporary literature, scientific writing has primarily been investigated as it is used in the research article. The RA is the most valuable currency in an industry which today has more publishers, journals, and researchers than ever before (Hyland & Jiang, 2019). Moreover, because of the dominance of English in academic publishing (Lillis & Curry, 2010), the RA has held unparalleled importance in the EAP literature (Samraj, 2016). Parts of the RA are often individually investigated, including abstracts, introductions, methods, discussions, findings, and author bios, among other parts. These scholars have examined not only scientific style, such as the grammar of noun phrases, but also citation practices, cohesion, communicative functions, cross-linguistic variation, disciplinarity, metadiscourse, multi-word expressions, stance and engagement, grammatical complexity, and voice and identity, among other features (see Samraj (2016) for some relevant citations). Thus researchers, especially in applied linguistics, have greatly expanded the linguistic and rhetorical features examined in RAs, revealing a rich and complex genre. For example, it is clearly incorrect to see all RAs (as thus all of scientific writing) as the same linguistic and rhetorical object. Writing across different disciplines often demands different rhetorical and linguistic organization (Biber & Gray, 2016; Hyland, 2007), even if the writers are academics writing for other academics.

In short, scientific writing fulfills unique communicative needs using features specifically developed for those needs. As some have highlighted, this fact has consequences for comparing professional and popular science writing. That is, despite their apparent similarity in topic and, in some cases, writer, popular science contrasts with professional science because the former responds to different situational demands and, as a result, likely relies on different rhetorical and linguistic features as well.

2.2.2 Discourse studies of popular science communication in natively digital genres

The digitization of science communication has generated increased interest from scholars looking to unpack how the Internet has transformed the way scientists and non-scientists communicate about academic research. Much of this literature adopts the lens of genre theory (Luzón and Pérez-Llantada, 2019), often with both pedagogical and sociological aims. Pedagogically, science communication has been an important area for fields within ESP because students often need communication skills necessary to succeed in academic environments. Sociologically, contexts of science communication have frequently been used to develop and apply theories of genre (e.g., Swales, 1990), which have mixed pedagogical and theoretical orientations (Bawarshi & Reiff, 2010). Thus, for these scholars, the increased digitization of science communication is fertile soil for examining and re-examining genre theory (Belcher, 2023; Pérez-Llantada, 2016).

Digital genres allow for the communication of science via written, spoken, or multi-modal modes. Among written forms, several scholars have investigated science blogs for their unique affordances (e.g., Hyland & Zou, 2020; Luzón, 2013, 2017; Sidler, 2016). Chief among these affordances is the quality of blogs to provide free-flowing interactive spaces where both experts and non-experts can share, critique, and comment on research (Blanchard, 2011). Other (primarily) written genres native to the Internet include science-focused crowdfunding projects (e.g., Mehlenbacher, 2017; Pérez-Llantada, 2021) and interactions on social media platforms like Twitter and Reddit (e.g., Moriarty & Mehlenbacher, 2019; Luzón, 2023), which provide both familiar and unique affordances for disseminating information. Other scholars have examined spoken and multimodal contexts of science communication, such as spoken podcasts (e.g., Ye, 2021), research videos (e.g., Luzón, 2019), graphical abstracts (e.g., Buehl, 2022), 3-Minute Thesis presentations (e.g., Liu, Tang, & Lim, 2023), and TED Talks (e.g., Valeiras-Jurado & Bernad-Mechó, 2022), among others.

Studies of natively digital genres have largely been interested in the texts' rhetorical features. One popular feature is organization, specifically the organization of chunks of text performing different functions known as rhetorical moves (e.g., Hu & Liu, 2018; Jiang & Qiu, 2022; Mehlenbacher, 2017; Ye, 2021). The results of many of these studies suggest that communicating scientific information, even across different contexts, demands attention to familiar details, such as literature review, research methodology, and study findings. But popularizers also make a greater effort to tailor information to and engage with non-expert audiences, embedding hyperlinks to additional explanations, ordering information to facilitate reader understanding, and personalizing the text with stance markers (Hyland & Zou, 2020; Luzón, 2013; Qiu & Jiang, 2021). Other studies have examined natively digital genres with more formally linguistic aims. For example, some Spanish-language crowdfunding proposals portray a hybridization of academic-style discourse, exemplified by elaborated noun phrases, with conversational discourse, exemplified by pronouns and modal verbs (Pérez-Llantada, 2021). Additionally, TED Talks have been shown to differ from university lecturers by speaking more quickly and using fewer academic words (Wingrove, 2017).

In short, these studies prove that technology has made science communication more abundant and influenced the very ways that people communicate it. A number of other studies examining more traditional genres of science communication elaborate on the linguistic features of popular science writing, which I turn to next in section 2.2.3.

2.2.3 Discourse studies of popular science writing in (digitally enhanced) traditional genres

While natively digital genres have proven to be exciting opportunities for examining new genres, scholars before and after the introduction of Web 2.0 technologies have explored the language of traditional genres used to communicate popular science, such as newspapers, magazines, and non-fiction books. As will be shown below, many of these texts also have digital versions, as the Internet has drastically impacted the printed periodicals industry. But unlike the genres described in section 2.2.2, digital genres like online newspapers and magazines transitioned into the digital age from their originally printed versions and thus portray similarities to their print counterparts.

Given the importance of linguistic description is this dissertation, a thorough summary of this section’s studies is essential. To that end, Table 2.1 provides a summary of 13 of the most relevant studies which empirically examined the language of popular science writing in print or digital periodicals. Studies which discussed but did not measure the linguistic features of their corpora (e.g., Adams-Smith, 1987; Hyland, 2010) are not shown in Table 2.1 but will be discussed following the table.

Table 2.1 An overview of corpus-based studies of the language of popular science writing in newspapers and magazines

Study	Corpus	Observations
Biber and Gray (2016)	**Corpus size**: larger* **Composition**: popularizations from academic journals **Topic**: mixed	• Popularizations use more verbs, adjectives, nominalizations, linking adverbials, and dependent clauses than research articles • Popularizations use fewer common nouns and prepositional phrases functioning as adverbials than research articles
Csongor and Rébék-Nagy (2012)	**Corpus size**: smaller **Composition**: popularizations from unknown websites **Topic**: health/medicine	• Popularizations make use of hedging language
Fu and Hyland (2014)	**Corpus size**: larger **Composition**: popularizations from online magazines **Topic**: mixed	• Popularizations construct an author-hidden discourse through limited expression of personal stance and use of hedging
Giannoni (2008)	**Corpus size**: medium **Composition**: editorials from academic journals **Topic**: medicine, linguistics	• Journal editorials make use of popularizing features, such as questions, metaphor, and personalization
Larsson (2019)	**Corpus size**: medium **Composition**: 20th century non-fiction books (BNC) **Topic**: mixed	• Popularizations mark a mid-way point between research articles and news reportage in its expression of stance

Larsson and Kaatari (2020)	**Corpus size**: medium **Composition**: 20th century non-fiction books (BNC) **Topic**: mixed	• Popularizations demonstrate noun complexity that is similar to news but more bare than research articles
Liu and Deng (2017)	**Corpus size**: larger **Composition**: popularizations from mixed online sources (COCA) **Topic**: science and technology	• Popularizations use more shell nouns than research articles, especially those expressing factual or modal meanings
Master (1991)	**Corpus size**: smaller **Composition**: popularizations from print magazines **Topic**: mixed	• Inanimate subjects in popularizations are more likely to be active voice than passive
Muñoz (2015)	**Corpus size**: larger **Composition**: popularizations from online magazines **Topic**: agriculture	• Semi-technical popularizations show an academic vocabulary profile between research articles and general news
Padula, Panza, and Muñoz (2020)	**Corpus size**: larger **Composition**: popularizations from online magazines **Topic**: engineering	• Semi-technical popularizations utilize sentence-initial *this* pronouns to create cohesion between sentences
Seoane and Suárez-Gómez (2020)	**Corpus size**: medium **Composition**: Excerpts of news and non-fiction texts written in Hong Kong English (ICE-HK) **Topic**: mixed	• Popularizations use fewer short passive voice verb phrases than academic writing but a similar amount of relative clauses

Varttala (1999)	**Corpus size**: medium **Composition**: popularizations from print magazines **Topic**: health/medicine	• Popularizations and academic writing use language expressing epistemic uncertainty to a similar extent
Zhang (2015)	**Corpus size**: medium **Composition**: Excerpts from non-fiction books written in British English (ICE-GB) **Topic**: mixed	• Popularizations use fewer extraposed clauses compared to academic writing, but also show a greater reliance on extraposed *to* clauses

*Corpus size: smaller (1-10 texts), medium (11-100 texts), larger (101+ texts)

The studies shown in Table 2.1 examined several genres of print and digital periodical, including multi-disciplinary academic journal articles, academic journal editorials, newspaper articles, magazine articles, and non-fiction books, all under the name of popular science. Most studies compiled private corpora of fewer than 100 texts characterizing discourse from several disciplines, though health and medicine was one of the more frequent domains. The focus of the studies ranged from metadiscoursal features to cohesion to vocabulary and more. Below I review findings from these and related studies.

Several scholars investigated features that reveal how writers interact with their readers and position themselves relative to the content. In particular, whether popular science writers are confident in their claims, in turn increasing the newsworthiness of their articles, or adopt a more tentative position has been examined, with somewhat mixed results. While both Hyland (2010) and Adams-Smith (1987) argue that popularizers avoid tentativeness to increase a story's newsworthiness, empirical evidence points in other directions. For example, Fu and Hyland (2014) conclude that popular science writers construct a cautious and author-hidden discourse style through minimal use of interactional features, such as hedges, boosters, and attitude markers, a position which is also adopted by Csongor and Rébék-Nagy's (2012) small-scale study of modal verbs in medical popularizations. Additionally, Larsson (2019) provided evidence that popular science writing may represent a mid-way point between the expression of stance in research articles and its expression in news reportage. Thus, while the motivations for avoiding tentative language in news writing are known, empirical evidence suggests a need for tentativeness when discussing scientific topics across audiences.

Another area of interest has been the way that popular science writers organize information from one sentence to another to be clear and coherent. Generally, scholars find differences in the use of cohesive devices between popular and professional science writing, with those differences being attributed to the need to tailor scientific information to diverse readers. For example, Nwogu and Bloor (1991) adopted a functional linguistic approach to analyze the thematic progression of sentences, finding that popularizations relied on a simple linear pattern where the predicate of one sentence becomes the subject of the next. The authors argue that this pattern is most useful for developing explanations. Padula, Panza, and Muñoz (2020) also find that popularizations make use of simple cohesive devices to meet the needs of its audience. Specifically, the researchers found a high rate of sentence-initial *this* pronouns functioning to link previous discourse to new discourse and guide the reader in its interpretation. This adoption of reader-friendly cohesive strategies can be thought of as "building bridges" between professional and popular discourse and has also been identified in the use of lexical cohesive devices in printed popularizations (Myers, 1991).

Several studies commented on the grammatical complexity of popular science writing, often from a register studies perspective (Biber & Gray, 2016; Larsson & Kaatari; 2020; Seoane & Suárez-Gómez, 2020). The general hypothesis states that popular science writing utilizes more verbs than nouns, more active than passive clauses, and noun phrases with less pre-modification (Adams-Smith, 1987), assertations that recent studies have largely supported. For example, Larsson and Kaatari (2020) observed their corpus of popular science writing to utilize greater noun phrase complexity compared to news writing but lesser complexity compared to research articles, reflecting the "mid-way point" finding in Larsson (2019). Additionally, Seoane and Suárez-Gómez (2020), in their comparison of academic writing and popular science writing in Hong Kong English, found academic writing to rely more heavily on the compressed short passive structure, while both corpora made similar use of the elaborated relative clause structure.

However, Biber and Gray's (2016) book-length treatment of grammatical complexity showed both support and rejection of the above-stated hypothesis. For example, the researchers found their popular science corpus to make greater use of verbs relative to nouns, as well as a greater use of dependent clauses in general, supporting the hypothesis. However, the researchers also found a greater use of adjectives in their popular science texts, which may signal greater noun premodification, as well as more nominalizations and passive voice verbs, countering the hypothesis. These mixed findings can likely be attributed to differences in the researchers' operationalization of popular science writing and the corpora compiled.

Remaining linguistic studies explored vocabulary (Muñoz, 2015), extraposed clauses (Zhang, 2015), and active verbs with inanimate subjects (Master, 1991), among other topics. Muñoz's (2015) study of the vocabulary of engineering popularizations suggested that these texts used academic words at a rate that places them in between research articles and general news, again pointing toward the mid-way point hypothesis described earlier. As a result, popularizations may serve as a useful pedagogical genre for teaching science writing to students who cannot or should not use research articles as their model. For example, Master (1991) argues that popularizations may be helpful texts for illustrating the use of inanimate subjects with active voice verbs, a prevalent feature in science prose but a potentially problematic one for students' whose first language does not permit such a pairing. Additionally, differences between popular and professional science writing may serve to raise students' awareness of the characteristics of the two genres (Parkinson & Adendorff, 2004). For example, while both genres make significant use of knowledge attribution, they do so in different ways and construct different perspectives on the scientific process (Calsamiglia & Ferrero, 2003; Parkinson & Adendorff, 2004).

As the above literature suggests, research of popular science discourse shows both common trends and methodological variation. While recent studies have examined in particular new digital genres, there is also a wide variety of datasets used to explore the language of popular science in written, spoken, and multi-modal modes. Below, in section 2.3, I highlight some gaps in this literature to situate the current dissertation among these studies.

2.3 Contributing to the literature on popular science discourse

In section 2.2, I reviewed previous work which contributed to the linguistic understanding of popular science discourse on local and global levels. In this section, I highlight four gaps in this literature to motivate this dissertation's study of popular science discourse.

First, the recent influx of research on science communication has primarily focused on digital multi-modal genres (e.g., Buehl, 2022; Jiang & Qiu, 2022; Liu, Tang, & Lim, 2023; Valeiras-Jurado & Bernad-Mechó, 2022; Xia, 2023), or on digital written genres in which the popularizer is the scientist (e.g., Hyland & Zou, 2020; Luzón, 2017; Sidler, 2016). TED Talk presentations, 3-Minute Thesis presentations, and science podcasts are exciting contexts that offer unique communicative affordances deserving of exploration, and personal blogs offer working scientists new contexts to present research narratives not possible in journal articles (Blanchard, 2011). However, the ever-increasing amount of academic research (Hyland, 2022) and benefits of digital publishing has also resulted in more science news reporting online, where once only found in print newspapers and magazines. These relatively unexplored publications are in the written mode and are produced by science journalists and freelancers. Thus, their characteristics represent both traditional print periodicals studied in earlier work and the new digital genres studied in recent years, and serve as resources of contemporary popular science writing produced by non-scientists.

Second, studies of digital genres like those cited above have primarily had rhetorical rather than linguistic aims. That is, they investigated features such as rhetorical moves (e.g., Hu & Liu, 2018; Jiang & Qiu, 2022; Mehlenbacher, 2017; Ye, 2021) and rhetorical strategies (e.g., Fahnestock, 1986; Giannoni, 2008; Hyland, 2010; Luzón, 2013), and multi-modal analyses investigated how the visual and/or auditory nature of the medium serves to communicate the scientific content of its speakers (e.g., Luzón, 2019; Valeiras-Jurado & Bernad-Mechó, 2022). As a result, linguistic analysis of digital genres remains comparatively unexplored (Peréz-Llantada, 2021). Further linguistic description of contemporary popular science writing may be particularly beneficial for pedagogical applications.

Third, most corpus-based studies of popular science writing published in print and digital periodicals do not take advantage of contemporary discourse or may not be generalizable. Studies which compiled personal corpora were often small due to technological limitations or the given purpose of the study. For example, several studies published in the previous century utilized corpora with about 15 or fewer texts (e.g., Adams-Smith, 1987; Garcés-Conejos & Sánchez-Macarro, 1998; Master, 1991; Myers, 1991; Nwogu, 1991; Vartalla, 1990). In addition, some recent, large-scale studies include older texts. Larsson and Kaatari (Larsson, 2019; Larsson & Kaatari, 2020) collected a subset of texts from the British National Corpus (BNC), including extracts from various non-fiction books published primarily in the 1980s. Zhang's (2015) popular science texts from the International Corpus of English-Great Britain (ICE-GB) includes extracts from various non-fiction books and magazines published in the early 1990s, while Seoane and Suárez-Gómez (2020) used texts from the Hong Kong component of the International Corpus of English-Hong Kong (ICE-HK). Other studies have investigated non-English language writing, primarily Spanish (e.g., Calsamiglia & Van Dijk, 2004; Peréz-

Llantada, 2021). As a result, few recent studies involve the compilation of contemporary English-language popular science writing published in online periodicals, and few corpora were designed to be generalizable.

Finally, corpus-based studies of popular science discourse tend to utilize a non-mixed methods approach, specifically textual analysis of their written, spoken, or multi-modal corpus of texts. As a result, there are few discourse analyses of popular science which include perspectives on composing texts from those who regularly do so. This is particularly lacking since the local and global features of popular texts are said to carry important communicative and rhetorical functions. Interviews, especially discourse-based interviews (e.g., Conrad, 2014; Hyland, 2007), can provide unique insight into the rationale behind why a text is composed the way that it is, as well as corroborate or challenge the researcher's intuition.

To summarize, the studies highlighted in section 2.1 and 2.2 explore the language of popular science discourse at different levels, including global levels like rhetorical moves and local levels like coverage of academic vocabulary. These studies have investigated popular science discourse communicated through new digital genres, such as online research videos and science-focused crowdfunding proposals, as well as traditional print and digital genres, such as newspaper and magazine columns. While these studies have contributed greatly to the growing interest in science communication directed at the public, there remain gaps in this literature. The current dissertation aims to fill some of these gaps by collecting a new corpus representative of contemporary online science news writing and investigating its discourse through linguistic analysis and discourse-based interviews. As a result, this dissertation attempts to contribute to the literature on popular science and science communication by providing linguistic insight into a corpus highly representative of contemporary popular science writing online.

3 METHODS

This dissertation adopts an explanatory sequential mixed methods design, wherein a qualitative method is used to help understand the findings of an initial quantitative analysis (Hashemi & Babaii, 2012). The initial quantitative study focuses on the analyses of a corpus of popular science writing and a corpus of matching RAs, which is described in section 3.1. The qualitative analysis includes discourse-based interviews with writers of popular science articles and is described in section 3.2.

3.1 The corpus study

3.1.1 Corpus collection

To collect a corpus of popular science writing in accordance with the gaps highlighted in section 2.3, I browsed the social media website Reddit and its forum “r/science,” which boasts over 20 million members, for sources of popular science commonly shared there. r/science users often post links to science articles and recent research to discuss with other users. Thus, r/science offers a good resource of contemporary popular science writing and discussion. I reviewed sources for those that publish articles written by science journalists about recently published research, a text type which one interview informant called “single-study articles,” because the articles are written primarily about a single recently published academic study. For easier reference, I will refer to these articles as science news articles, or SNAs.

From the review of r/science, I selected eight sources from which to collect SNAs. Review of these sources suggested that four topical domains were most common among popular science writing generally, namely health/medicine, mental health/psychology, space/physics, and the environment. From each of these four domains, I collected 100 SNAs, 50 from one source and 50 from another source, resulting in a final corpus of 400 SNAs across four topical domains and eight sources. All SNAs were authored by a single writer and linked to the source of the academic research cited in the article.

SNAs were collected both automatically and manually. Initially, larger corpora for most sources were collected using the package *rvest* (Wickham, 2021) in R (R Core Team, 2020). SNAs were then identified by searching for key phrases such as *a recently published article* or *recently published in* and inspecting those texts. However, SNAs from one source were manually downloaded from its website due to difficulty in automatic collection. I established a ceiling of 10 SNAs that any single author could contribute to the corpus, meaning that no author represented >2.5% of all SNAs. A description of the SNA corpus is shown in Table 3.1.

Table 3.1 A description of the SNA corpus

Category	Source	# of Authors	Date Range	# of Texts	# of Words
Health/ medicine	NewAtlas	9	2018-2021	50	23,842
	ScienceNews	20	2019-2021	50	32,871
Psychology/ mental health	PsyPost	6	2017-2021	50	24,304
	The Academic Times	14	2019-2021	50	45,153
Space/ physics	Astronomy	19	2018-2021	50	41,592
	Inverse	21	2018-2021	50	44,431
Environment	Science Alert	8	2017-2021	50	33,128
	ZME Science	10	2018-2021	50	36,141
Total		107	2017-2021	400	281,444

Table 3.1 shows that the SNA corpus is comprised of 400 texts equating to 281,444 words written by 107 unique authors across four years.

Since each SNA reviewed a recently published academic article, a matching corpus of 400 RAs was also collected. After manually identifying the primary RA reviewed by a given SNA, I searched for and downloaded that RA; converted it into a Microsoft Word document to remove tables, figures, references, abstracts, and formulae; converted the Word document into a text file; and read the text file into R for cleaning and analysis. A description of the RA corpus is shown in Table 3.2.

Table 3.2 A description of the RA corpus

Category	Total Sources (5 Most Frequent)	Date Range	# of Texts	# of Words
Health	47 journals 1. Science Translation Medicine 2. Nature 3. Nature Medicine 4. Science 5. The New England Journal of Medicine	2018-2021	100	458,714
Psychology	79 journals 1. Frontiers in Psychology 2. Personality and Individual Differences 3. Psychological Medicine 4. American Behavioral Scientist 5. Journal of Organizational Behavior	2017-2021	100	597,085
Space	33 journals 1. Nature Astronomy 2. Science 3. The Astrophysical Journal Letters 4. The Astrophysical Journal 5. Proceedings of the National Academy of Science of the United States of America	2017-2021	100	485,919
Environment	81 journals 1. Nature 2. Nature Communications 3. Scientific Reports 4. Nature Climate Change 5. Nature Geoscience	2017-2021	100	472,572
Total	240 journals	2017-2021	400	2,014,290

Table 3.2 shows that the RA corpus is comprised of 400 texts equating to 2,014,290 words published in 240 different journals across four years.

3.1.2 Corpus evaluation

Since the goal of this dissertation is to provide generalizable linguistic insight into SNAs and RAs, it is important to consider to what extent the corpora are representative of their target domains. I take corpus representativeness to consist of two factors: domain considerations and distribution considerations (Egbert, Biber, & Gray, 2022). Domain considerations include qualitative decisions about what should be included in a corpus based on a description of the target domain. Distribution considerations refer to how many texts should be collected to allow for accurate and generalizable linguistic counts. Below, I evaluate the SNA corpus and RA corpus based on these two pillars of corpus representativeness.

Appendix A provides a detailed description of the domain considerations for the SNA corpus. Here, I describe that description in brief. The SNA corpus is representative of a specific sub-register of popular science, namely single-study online news articles. The target domain was clearly operationalized when collecting texts, leading to strong internal consistency within the corpus. For example, an inconsistent corpus of science news writing might include news reports, photo and video articles, editorials, and book reviews. However, this dissertation's SNA corpus includes only whole texts of about 700 words of length written by one science writer about a recently published research article. Hence, they are likely to share similar situational characteristics (more on the situational characteristics in Chapter 4).

On the other hand, one limitation is in the non-randomness with which the sources and texts were selected. Sources were identified by a manual review of posts shared on r/science, but these sources do not necessarily represent the most frequently shared sources on r/science or in general. Moreover, I did not pay due attention to the month and year that each article was published in, leading to certain dates being more represented than others. Similarly, the RA corpus is even less balanced, owing to the fact that these texts were collected strictly on the criterion that they were reviewed by an SNA. Thus, certain journals like *Nature* and *Science* are much more highly represented than others, and year of publication is directly related to the year of publication of the corresponding SNA.

Distribution considerations often start and end with the number of texts or words in a corpus. In this sense, the SNA and RA corpora are neither 'small' nor 'large' by contemporary (albeit subjective) standards. Both corpora have 400 texts, but the SNA corpus has about 1/7th the number of words that the RA corpus does, owing to its much shorter texts. However, a more precise way of determining the appropriateness of these sizes is relative standard error (RSE), proposed by Egbert et al. (2022). RSE is calculated as follows:

$$RSE = \frac{standard\ error\ (SE)}{mean\ rate\ of\ occurrence}$$

where,

$$SE = \frac{sample\ standard\ deviation}{\sqrt{sample\ size}}$$

In combination with a chosen critical value, RSE describes the precision with which a corpus can provide accurate measurements of linguistic features. Table 3.3 presents three critical RSE values associated with a given error rate (see Egbert et al., 2022, p. 138 for how these critical values were calculated).

Table 3.3 Critical RSE values for several error rates

Error rate (as a percentage of the mean)	RSE
10%	.0510
5%	.0255
1%	.0051

An RSE of 0.0255 for a given linguistic feature means that the true mean for that feature in the target domain is plus or minus 5% of the corpus's observed mean. So, if progressive aspect verbs occur 10 times per 1,000 words (ptw) in a spoken English corpus with an RSE of 0.0255, then one can conclude that the true mean is within 9.5 and 10.5 ptw. In short, a corpus which provides RSE values close to 0.0051 is highly generalizable to the target domain, while a corpus with values that near 0.0510 and above shows room for improvement.

RSE values for a handful of relevant linguistic features in the SNA and RA corpora are shown in Table 3.4.

Table 3.3 RSE values for several linguistic features across the registers

Linguistic features	RSE	
	RA corpus	SNA corpus
All verbs	0.0063	0.0057
Past tense verbs	0.0287	0.0265
Perfect aspect verbs	0.0319	0.0337
Central modal verbs	0.0248	0.0251
Passive voice verbs	0.0248	0.0251
Transitive verb patterns	0.0153	0.0192
Copular verb patterns	0.0188	0.0309

Table 3.4 shows that, in general, the RA corpus provides more precise measurements of linguistic features, owing to their longer articles and smaller standard deviations. However, the SNA corpus, despite their much shorter texts, shows similar RSE values for most features shown in Table 3.4. While these features are not comprehensive, they suggest that the SNA corpus is capable of providing accurate, generalizable measurements of features relevant to this study.

3.1.3 Feature selection

As explained in section 2.3, many recent studies have investigated the *rhetorical* features of popular science discourse. This dissertation contributes to that literature by focusing on the *linguistic* description of popular science. Adopting the perspective of register analysis, which states that linguistic features have functional purposes in discourse (Biber & Conrad, 2019), I selected features of the short and long verb phrase to compare across the SNA and RA registers. Specifically, this dissertation investigates variation in the short verb phrase, or the ways that verbs vary for tense, aspect, voice, and modality, as well as variation in the long verb phrase, or the ways that elements of the clause are arranged to constitute 'verb patterns' (also called valency patterns). Verb patterns include core elements such as subjects, verbs, and objects, as well as optional adverbials.

There are at least three good reasons why a focus on the verb phrase is insightful. First, the verb holds a privileged position among the word classes in that it partially controls the other elements of the clause (Biber et al., 1999), such as whether a phrasal object, clausal object, or no object follows the verb. Additionally, the verb varies in grammatical form more so than other word classes, including for tense, aspect, voice, and modality, each of which has rich functional relationships in discourse. Second, early genre research of scientific writing also found use in studying the verb. Trimble (1985), for example, dedicates a chapter to describing the rhetorical-grammatical relationships of tense, voice, and modal verbs in English for Science and Technology (EST) writing. Finally, a focus on discrete linguistic features can be useful for creating pedagogical materials.

In section 3.1.3.1, I describe the features of the long verb phrase that will be examined in this dissertation. Then, in section 3.1.3.2, I describe features of the short verb phrase that will be examined. Finally, I describe an analysis of reporting clauses in section 3.1.3.3.

3.1.3.1 Verb patterns of the main clause

Verbs partially control the other elements that appear in their clauses (Biber et al., 1999). Thus, some sentences require a subject, verb, and direct object, while others only a verb. The patterning of these elements is traditionally referred to as valency (Allerton, 1982), with some well-known patterns including copular, monotransitive, and complex transitive patterns (Quirk et al., 1985). Valency patterns and other similar concepts have been examined for the purposes of teaching English (e.g., Francis et al., 1996; Ma & Qian, 2020), tracking second language acquisition (e.g., Ellis, Römer, & O’Donnell, 2016), dictionary creation (e.g., Herbst et al., 2004), and genre comparison (e.g., Casal, Shirai, & Lu, 2022).

In this dissertation, I use the term “verb pattern” as a theory-neutral term to refer to many of the characteristics shared by descriptions of valency patterns. However, the descriptions of verb patterns in this dissertation often borrow from Biber et al. (1999), referring to such familiar elements as complement clauses, direct objects, optional adverbials, and so on. Importantly, this dissertation only considers the verb patterns and their obligatory and optional elements of main, independent clauses. Thus, the verb patterns of dependent adverbial clauses, for example, are not included in the analysis.

3.1.3.2 Variation in the short verb phrase

Finite verbs can vary for tense, aspect, modality, and voice, all of which have functional purposes in discourse. For example, the present tense can be used to emphasize the generalizability of research findings and present perfect to show disagreement with past research (Salager-Mayer, 1992), while the simple past and passive voice are frequent in the methods sections of research articles (Heslot, 1982). The passive voice can be used to indicate that a method is standard procedure or to refer to others' research (Tarone, Dwyer, Gillette, & Icke, 1998). Modal verbs express writer stance (Biber et al., 1999) and can be used to mark research hypotheses (De Waard & Pander Maat, 2012). Additionally, both the density of verbs and their grammar vary as a function of register (Biber et al., 1999), and thus can be used to reveal register differences.

In this dissertation, a distinction is first made between finite and non-finite verb phrases. A finite verb phrase includes a lexical verb and any auxiliary and can be either tensed or modalized. By contrast, non-finite verbs are neither tensed nor modalized. With regard to finite verb phrases, this dissertation considers the tense (past or present), aspect (simple, perfect, or progressive), voice (active or passive), and modality (central and semi- modals) of verbs. With regard to finite modal verb phrases, a distinction is made between central modals and semi-modals. While central modal verbs are a closed word class typically including nine verbs, the class of semi-modal verbs is more porous (see, e.g., Leech et al., 2009; Palmer, 1990). This dissertation includes only the semi-modals found to be moderately frequent in news or academic writing, namely *be going to*, *have to*, and *need to* (Biber et al., 1999, p. 489). With regard to non-finite verb phrases, this dissertation considers both the aspect and voice of these verbs.

3.1.3.3 Reporting clauses

Diverging somewhat from the features discussed in 3.1.3.1 and 3.1.3.2, reporting clauses are clauses that report the direct or indirect speech or thought of others (Biber et al., 1999). Attributing words or thoughts to others is an essential feature of both academic and popular science writing. In the context of popular science writing, reporting clauses often include direct quotations from important actors in a story (Calsamiglia & Ferrero, 2003), while academic writers typically utilize in-text citation methods (Thompson & Ye, 1991). Reporting clauses vary in their choice of verb and the grammar of that verb, which are linked to functional purposes like expressing agreement or tentativeness (e.g., Salagar-Mayer, 1992; Thompson & Ye, 1991).

Reporting clauses are often found in the form of *that* complement clauses following the verb (e.g., *Skinner found that…*) or anticipatory *it* clauses in the passive voice (e.g., *It has been suggested that…*) (Charles, 2006; Francis et al., 1996). To these two patterns I also add sentences including the prepositional phrase *according to* + noun phrase (e.g., *..., according to the study's author*). I will refer to these patterns using the following shorthand: TV+*that.*CL, *it*+*BE*+TV+*that.*CL, and *according+to*, where 'TV' refers to 'transitive verb' and '*that.*CL' to a *that* complement clause. These three patterns are useful for the current dissertation because they do not begin by identifying in-text citations and then move to analyzing its surrounding discourse, which would bias the RA corpus due to its use of citation, but instead by identifying clauses and sentences that typically function to attribute words and ideas. In other words, they can identify attribution whether there is explicit use of citation or not.

3.1.4 Identification and accuracy of linguistic features

To identify the linguistic features described in section 3.1.3, Python scripts (Python Software Foundation, 2021) were written. The package *spaCy* (v.3) (spaCy, 2022) was used to part-of-speech tag and syntactically parse the corpora. *spaCy* is a fast and accurate NLP tool trained on the *OntoNotes* corpus, reporting 97.8% tagging accuracy and 95.1% parsing accuracy there (Facts & Figures, n.d.). The scripts tag and parse corpus texts and then write the retrieved information onto text files, which can then be read into R for cleaning and statistical analysis. In section 3.1.4.1, I provide descriptions of these scripts and their accuracy rates. For accuracy rates, an assistant with experience with descriptive grammar was trained to review samples of data retrieved for verb pattern analysis, short verb phrase variation, and reporting clauses. More detail is given below.

3.1.4.1 Identification and accuracy of verb patterns

To identify verb patterns, a script was written to break each text into main clauses (typically at sentence boundaries). Then, the main verb of the clause is identified, which allows for the other elements of the verb pattern to also be identified and examined. The script works sentence by sentence to return this information for each text. An example output for one clause is shown below:

```
@VERB:examine @PhV:none @PrV:none @SMV:none
@PAT:Aadvcl(16)S(2)@SUBJECT:study@VDONOUN(25)@DO:impact@SEN: While our
inclusion criteria and design of our experiment helped maximize
participant comprehension and testing stamina, future studies could
examine the impact of music on emotions in individuals at more advanced
stages of the disease as well as with other neurodegenerative diseases
(e.g., frontotemporal dementia).
```

This output shares, in order, that the main verb lemma (VERB) is *examine*; there are no phrasal (PhV), prepositional (PrV), or semi-modal (SMV) verbs; the verb pattern (PAT) begins with an adverbial clause of 16 words (Aadvcl(16)), followed by a subject of 2 words (S(2)) with *study* as its head (SUBJECT), the main verb (V), and a singular noun direct object (DONOUN) with a length of 25 words and *impact* as its head. Finally, the full sentence (SEN) is returned.

Hundreds of unique verb patterns were identified by the program, most of which occurred very rarely. After careful review, the top 30 most frequent patterns were selected for possible analysis. These patterns represented 91% of all clauses identified by the program in the SNA corpus and 93% in the RA corpus. Table 3.5 describes these 30 verb patterns.

Table 3.4 The 30 most frequent verb patterns across the corpora (in alphabetical order)

Verb pattern	Description	Example
SV	Subj + lexical verb	More answers may come from upcoming missions.
SV?	Subj + lexical verb + ?	Why do these stars come together?
SVc(adj)	Subj + copular verb + adjective phrase complement	This negative relationship was significant in 20 countries, though the effect was stronger in WEIRD countries than non-WEIRD countries.
SVc(adj)that.CL	Extraposed *it* clause with *that* complement clause	Considering how common asthma is, it's rather surprising that we don't know all that much about how the condition works.
SVc(adj)to.INF	Extraposed *it* clause with *to* infinitive clause	By analyzing these dust layers, it's possible to infer many qualities about the climate they originate from.
SVc(adp)	Subj + copular verb + prepositional phrase complement	This is in line with previous estimates (and even a bit more conservative than some).
SVc(noun)	Subj + copular verb + noun phrase complement	“That’s a recipe for disaster.”
SVc(pron)	Subj + copular verb + noun phrase headed by pronoun complement	They're some of the few objects that can be seen from such a distance.
SVcthat.CL	Subj + copular verb + *that* clause complement	The idea is that if the immunotherapy starts to get out of control, doctors could administer lenalidomide to switch off the rogue CAR T cells.
SVcto.INF	Subj + copular verb + *to* infinitive clause complement	The goal is to identify conditions in which the rate of destruction is faster than the rate of production.
SVcwh.CL	Subj + copular verb + *wh* clause complement	That's what the team in Colorado set out to do.
SVDO(noun)	Subj + lexical verb + noun phrase functioning as direct object	So the researchers projected each respondent's political ad viewing separately.

SVDO(noun)to.INF	Subj + lexical verb + noun phrase functioning as direct object + *to* infinitive clause complement	Bowman and her colleagues recruited 142 pairs of mothers and infants to participate in their experiment.
SVDO(pron)	Subj + lexical verb + noun phrase headed by pronoun functioning as direct object	One of the first interventions to amass considerable research support used nothing more than some cans of colorful paint.
SVDO(pron)OP(adj)	Subj + lexical verb + noun phrase headed by pronoun functioning as direct object + adjective phrase functioning as object predicative	This, in turn, makes people more likely to ethically condemn such articles and to share them on social media.
SVDO(propn)	Subj + lexical verb + noun phrase headed by proper noun functioning as direct object	The survey included the Trait Emotional Intelligence Questionnaire.
SVIO(pron)DO(noun)	Subj + lexical verb + noun phrase headed by pronoun functioning as indirect object + noun phrase functioning as direct object	"Then that gives us a sample that we can start to study."
SVPass	Subj + passive verb	The findings are reported in Nature.
SVPassthat.CL	Extraposed *it* passive clause with *that* complement clause	It has been acknowledged that childhood traits and circumstances like lower intelligence or childhood deprivation can have a negative impact on health and morbidity later in life.
SVPassto.INF	Subj + passive verb + *to* infinitive complement clause	Hypatia is thought to be a fragment of an extraterrestrial rock originally several meters long, which segmented into numerous pieces during its journey to Earth.
SVthat.CL	Subj + lexical verb + *that* complement clause	However, now we see that some corals will also ingest plastic because of chemicals leaching from the microplastic particles.

SVto.INF	Subj + lexical verb + *to* infinitive complement clause	They also fail to provide any sensory feedback.
SVV.ing	Subj + lexical verb + *ing* complement clause	This step involved chopping up DNA from a bacterium of interest and inserting individual snippets into E. coli cells.
SVwh.CL	Subj + lexical verb + *wh* complement clause	Common factors considered include whether liquid water could exist on the world.
that.CLSV	*that* complement clause + subj + lexical verb	"That it could have taken 15 million years is a good thing to know," he says.
that.CLSVDO(propn)	*that* complement clause + subj + lexical verb + noun phrase headed by proper noun functioning as direct object	People are "more uncomfortable with it [incivility] from women […]," Bauer told The Academic Times.
that.CLVS	*that* complement clause + lexical verb + subj	"Very few marine organisms stayed in the same habitats they were living in […]," said oceanographer Curtis Deutsch of the University of Washington.
ThereVc(noun)	Existential *there* + copular verb + noun phrase complement	First of all, there already are some drugs which are delivered via tablets that the patient places beneath their tongue.
V.ingVc(adj)	*ing* complement clause + copular verb + adjective phrase complement	But studying muons is difficult.
V.ingVDO(noun)	*ing* complement clause + lexical verb + noun phrase functioning as direct object	Disabling or removing that pump stopped bacteria from developing resistance.

Intransitive patterns. Just one pattern in Table 3.5, namely SV, represents an intransitive verb pattern, or a pattern requiring only a lexical verb and subject. However, accurately identifying intransitive patterns was particularly challenging for the program. This was in part due to the presence of multi-word verbs, which the program would often mistake for an intransitive verb followed by an optional adverbial. Thus, I reviewed the most common verbs identified as intransitives to distinguish those actually functioning as intransitives from those that were not, resulting in between 40-48 intransitive verbs in either corpus. Appendix B lists the intransitive verbs retained in the analyses.

Transitive patterns with phrasal objects. Several patterns in Table 3.5 represent transitive verb patterns with phrasal objects. Most of these patterns are monotransitive (e.g., SVDO(noun), SVDO(pron)), including just one phrasal object, while one pattern includes a direct and indirect phrasal object (SVIO(pron)DO(noun)). Another variation includes a phrasal object and object predicative (SVDO(pron)OP(adj)), and finally one pattern has a non-finite clause as subject and phrasal object following the verb (V.ingVDO(noun)). To accurately identify those multi-word verbs functionting as transitive verbs rather than intransitive verbs followed by an adverbial, I gathered 269 of the most frequent phrasal verbs and prepositional verbs from Biber et al. (1990, Ch. 5) and Garnier and Schmitt (2014) and used these to identify the most common multi-word transitive verbs in the corpora. Appendix C lists these multi-word verbs retained in the analyses.

Transitive patterns with clausal complements. A large number of patterns in Table 3.5 include a transitive verb with a finite or non-finite complement clause. Most of these patterns have a traditional syntactic order, where the subject and verb precede the complement clause (e.g., SVthat.CL); however, several patterns where the clause precedes the subject and verb were also attested (e.g., that.CLSV). Transitive patterns could include either finite complement clauses (e.g., SVthat.CL, SVwh.CL) or non-finite complement clauses (e.g., SVto.INF, SVV.ing). In some cases, the subject itself was a clause, as shown earlier.

Transitive patterns with passive voice. Several verb patterns in Table 3.5 include transitive verbs with phrasal or clausal complements, but their verbs are in the passive rather than active voice. Specifically, three verb patterns have a passive transitive verb, namely SVPass, SVPassthat.CL, and SVPassto.INF. As these patterns illustrate, SVPass includes no other complement other than the subject, while SVPassthat.CL and SVPassto.INF both include a complement clause.

Patterns with copular verbs. Several patterns in Table 3.5 include a copular verb, or a verb which links phrases and clauses to their subjects, either characterizing them or locating them in time and space (Biber et al., 1999). These patterns always included a subject and a copular verb (often the *be* verb), while the phrase or clause linked to the subject could be an adjective phrase (SVc(adj)), noun phrase (SVc(noun)), prepositional phrase (SVc(adp)), pronoun (SVc(pron)), or clause (e.g., SVcthat.CL). Two patterns involved an extraposed clause, leading to three clausal elements (SVc(adj)to.INF, SVc(adj)that.CL).

Other patterns. The final two patterns include interrogative clauses and existential *there* clauses. The SV? pattern refers to interrogative clauses. While some of these patterns were intransitive, others included a phrasal or clausal object. In all cases, the sentence had the syntax of *wh*-word or yes/no questions. ThereVc(noun) refers to existential *there* patterns, which are comprised of a semantically empty subject, *there*, followed by a form of the *BE* verb and a noun phrase complement. Such patterns signify that the noun phrase exists in time and space.

To assess the script’s accuracy, I selected up to 40 random sentences from each pattern, 20 from either corpus, resulting in a review of 1141 sentences. I found all but one pattern to have an accuracy of 90% or above, namely SVwh.CL. This pattern showed low accuracy in the RA corpus specifically (15/20 or 75%), where non-finite complement clauses were sometimes misidentified as finite *wh*-word clauses. Nonetheless, I considered this percentage and the rest as sufficiently high to continue with the analyses.

3.1.4.2 Identification and accuracy of short verb phrase variation

A different script serves to identify variation in the short verb phrase by splitting texts into sentences and searching for words tagged as lexical or auxiliary verbs, which are then examined for their finiteness, tense, voice, and aspect. An example output from this script is shown below:

```
@type:nf_ing@verb:plan@SMV:@NEG:@PASS:@1@SEN:Despite "green" burials becoming
increasingly available in North America, some older eco-conscious adults remain unaware of the
option when planning for their deaths, a small study hints.
```

This output shares, in order, that a non-finite -*ing* clause (type:nf_ing) whose lemma is *plan* (verb:plan) was identified; that there is no semi-modal verb (SMV:); that the verb is assertive (NEG:) and active voice (PASS:); and the verb was found in the first text of the corpus (1). Finally, the full sentence is returned (SEN:).

To assess the accuracy of the script, a random sample of up to 60 sentences (30 from each register) of all possible tense-aspect-modality combinations were reviewed by a trained assistant. So, for example, 60 sentences containing -*ing* clauses with simple aspect were reviewed, 60 sentences containing modal verb phrases with perfect aspect were reviewed, 60 sentences containing past tense verbs with simple aspect were reviewed, and so on. In the case that there were not 60 different sentences for a given grammatical combination, the full number of sentences was reviewed. Table 3.6 reports on the accuracy of 14 different tense-aspect-modality (TAM) combinations, as well as their voice.

Table 3.5 Accuracy rates for tense, aspect, modality, and voice of verbs

Variation*	N	TAM %	Voice %
Modal perfect	59	98%	100%
Modal simple	60	97%	100%
Past perfect	52	94%	100%
Past perfect progressive	17	100%	100%
Past progressive	54	98%	100%
Present perfect	60	97%	100%
Present perfect progressive	42	98%	100%
Present progressive	57	91%	100%
Present simple	60	100%	100%
Simple bare infinitive	56	98%	100%
Simple infinitive	59	98%	100%
Perfect infinitive	42	88%	90%
Simple *-ing* participle	60	100%	100%
Simple *-ed* participle	60	97%	97%
Total	738	97%	99%

Note: Only six perfect aspect *-ing* participles were found and only half were accurate, so these were dropped from the analysis. Any combinations not shown here were not attested for.

Table 3.6 reports a 97% accuracy with regard to tense, aspect, and modality and 99% accuracy with regard to voice on 738 verbs. I deemed these rates to be acceptable.

3.1.4.3 Identification and accuracy of reporting clauses

Another script was written to identify reporting clauses. The script splits texts into sentences and identifies their main verbs. Then, it identifies whether a TV+*that.*CL pattern, *it*+*be*+TV+*that.*CL pattern, or an *according to* prepositional phrase is present in the sentence. An example output from this script is shown below:

```
@tv+that:TRUE@subordinator:that@subject:they@verb:suggest@255@SEN:Instead, they
suggest that Venus' phosphine may have made its way to the planet's upper atmosphere by
way of volcanic eruptions.
```

This output shares, in order, that a TV+*that*CL pattern is present (tv+that:TRUE); it has an explicit subordinator (subordinator:that); its subject lemma is *they* (subject:they); its main verb lemma is *suggest* (verb:suggest); and the sentence was found in text ID number 255. Finally, the full sentence is returned (SEN:).

A random sample of 20% of each of the three patterns was extracted and analyzed for its accuracy. Regarding TV+*that.*CL, 93% accuracy on 326 instances was reported for the SNA corpus, and 94% on 1,431 instances was reported for the RA corpus. However, because there were fewer instances of the *it*+*be*+TV+*that.*CL (18 in SNA, 169 in RA) and *according to* (201 in SNA, 161 in RA) patterns in the corpora, all instances were reviewed and found to be 100% accurate.

Section 7.1 of Chapter 7 describes four codes by which reporting clauses were categorized for functional analysis. However, it is necessary to note here a methodological step taken to apply those codes. Because of the large number of TV+*that*.CL patterns in the corpora (over 10,000 instances between the registers), it was not plausible to code all instances. Instead, to provide an accurate estimation of the rate of TV+*that*CL patterns functioning as reporting clauses, the following steps were taken.

First, all instances of TV+*that*CL patterns in the SNA corpus were printed to a .csv file, which was then sorted by the source of the sentence (i.e., which of the eight SNA outlets the sentence originated from). Then, each instance was coded for whether it functioned as a reporting clause or not until each source had 100 instances of TV+*that*CL patterns functioning as reporting clauses, resulting in 800 instances. The same process was applied to the RA corpus; however, instead of sorting by source (since there were over 100 unique journals), all TV+*that*CL patterns were placed in a random order and coded until 800 reporting clauses were identified.

As a result of this process, both corpora had a comparable pool of TV+*that*CL patterns functioning both as reporting clauses and not as reporting clauses. The estimation of the rate of TV+*that*CL reporting clauses was then calculated as follows:

$$Normalized\ rate = \frac{X}{\#\ of\ words\ in\ corpus} \times 1000$$

where,

$$X = Total\ \#\ of\ instances \times \left(\frac{\#\ of\ instances\ functioning\ as\ RCs}{\#\ of\ instances\ not\ functioning\ as\ RCs}\right)$$

In short, this formula multiplies the proportion of reporting clauses to non-reporting clauses of a sample with the larger population, which is then multiplied by a normalization rate, to produce an estimate of the number of TV+*that*CL reporting clauses among all *that* complement clauses in the corpora.

3.1.5 Statistical analysis of linguistic features

In this section, I describe the statistical tests applied to the analysis of verb patterns, short verb phrase variation, and reporting clauses, including calculating normalized rates (3.1.5.1), comparing means with confidence intervals (3.1.5.2), comparing categorical data (3.1.5.3), linear mixed effects models (3.1.5.4), and cluster analysis (3.1.5.5).

3.1.5.1 Comparing frequencies across the corpora

Given the different sizes of the corpora, I present the normalized rates of occurrence for all features rather than raw counts, unless otherwise noted. Excluding reporting clauses (see section 3.1.4.3), normalized rates were produced per text rather than per corpus, allowing me to perform inferential statistical tests on their values. Normalized rates of occurrence are calculated as follows:

$$Normalized\ rate\ of\ occurrence = \frac{Raw\ frequency\ of\ feature\ X\ in\ Text\ Y}{Total\ number\ of\ words\ in\ Text\ Y} \times 1{,}000$$

As an example, if 7 modal verbs occurred in an SNA 600 words in length, the normalized frequency of modal verbs in that text would be 11.67 per one thousand words (ptw).

3.1.5.2 Testing differences with numeric variables

To test the hypothesis whether two means are drawn from the same population, I compare the overlap or non-overlap of 95% confidence intervals (CIs), as opposed to typical statistical tests like the *t*-test, which produces p values. P values are dichotomous and not highly informative on their own, while 95% CIs visualize the probability of rejecting the null hypothesis as well as the strength of that rejection. Thus, they are more intuitive and easier to interpret than p values alone (Wallis, 2021). In this dissertation, 95% CIs are calculated using the following formula (from Cumming et al., 2007):

$$95\%\, CIs = mean \pm\ SE *\ t(n-1)$$

where,

$$t(n-1) \approx 1.96$$

Put another way, 95% CIs are computed by multiplying the standard error of the mean by a critical value of the student's *t* distribution, determined by the degrees of freedom of a given test (i.e., n-1) and the desired level of confidence. For this dissertation, this critical value equates to about 1.96, which matches the degrees of freedom of 399 (i.e., 400 texts – 1 text = 399 texts, or degrees of freedom) and the level of confidence is 95%, or 0.95, on a table of critical *t* values.

After adding and subtracting the 95% CI from a given feature's mean, a range of values is produced, which can be compared with the range of another sample to infer whether they are different. If the ranges overlap, there is not enough evidence to conclude that they are significantly different. If they do *not* overlap, there is evidence for a significant difference (Brezina, 2018). Schenker and Gentleman (2001) note that this method can fail to detect significant differences where they in fact exist, but the sample sizes in this study are large enough to be unconcerned.

To examine whether a significant difference is meaningful, I adopt Cohen's *d* as an effect size, calculated using the following formula (from Cohen, 1988):

$$d = \frac{M1 - M2}{\sqrt{\frac{SD1 + SD2}{2}}}$$

The values 0.2, 0.5, and 0.8 correspond to small, medium, and large effect sizes, respectively (Cohen, 1988).

3.1.5.3 Testing differences with categorical variables

Confidence intervals were used when a numeric variable was compared across the registers, but in some cases I compare the distribution of a categorical variable instead. In such cases, a Chi-square test is most appropriate. The following formula was used to calculate the Chi-square statistic (from Sheskin, 2011, p. 280):

$$\chi^2 = \sum_{i=1}^{k}[\frac{(O_i - E_i)^2}{E_i}]$$

Put another way, the Chi-square statistic is calculated by subtracting the observed frequency by the expected frequency, squaring that value, and dividing it by the expected frequency. This is done for each cell in the contingency table and the resulting values added together to net χ^2. In this dissertation, the frequencies plugged into Chi-square tests were normalized rather than raw to account for differences in sample sizes.

In a Chi-square goodness-of-fit test, there is only one categorical variable, and the expected value is determined *a priori* by the researcher. In a Chi-square test of independence, there are two independent samples, and the expected value is calculated by multiplying the row total with the column total of a given cell and dividing that value by the total sample size. Both kinds of tests were used in this dissertation.

Because Chi-square tests are omnibus tests, the source of any statistical significance is not clear (Sharpe, 2015). However, the source can be identified by observing the residuals of each cell, or the differences between the observed and expected values (Delucchi, 1993). Because cells with larger frequencies will generally result in larger residuals, raw residuals are normalized by dividing the difference by the square root of the expected frequency. Such standardized residuals give a better comparison across cells, with those larger than ± 2 indicating significant differences (Agresti, 2007).

Finally, to measure the meaningfulness of a significant χ^2, Cramer's *V* effect size was adopted. Cramer's *V* is calculated as follows (from Brezina, 2018, p. 114):

$$Cramer's\ V = \sqrt{\frac{\chi^2}{total\ observations \times (\#\ of rows\ or\ columns, whichever\ is\ smaller - 1)}}$$

Interpretation of the magnitude of Cramer's *V* depends on the degrees of freedom in the contingency table, as shown in Table 3.7 (from Brezina, 2018, p. 115).

Table 3.6 Interpretations for Cramer's V *effect sizes*

Degrees of freedom	Effect size		
	Small	Medium	Large
1 (2×2)	.10	.30	.50
2 (2×3 or 3×2)	.07	.21	.35
3 (2×4 or 4×2)	.06	.17	.29

For example, a Cramer's *V* of 0.35 on a contingency table with two rows and two columns would correspond to a medium effect size.

Finally, in some circumstances, I examine the difference between two categorical variables but do not test the statistical significance of that difference. In such a case, the odds ratio, which compares the chances of one outcome over the chances of another, can serve as a straightforward measure of effect size (Levshina, 2015). Odds ratio is calculated as follows:

$$Odds\ ratio = \frac{Odds\ of\ outcome\ X}{Odds\ of\ outcome\ Y}$$

where,

$$Odds = \frac{Successful\ outcome}{Unsuccessful\ outcome}$$

If both outcomes are the same, the odds ratio will be 1. The greater the odds of X, the higher the value will be above 1, while the greater the odds of Y, the closer the value will be to 0.

3.1.5.4 Testing interactions with linear mixed effects models

In some cases, I adopt a statistical test known a linear mixed effects (LME) model. A mixed model is a powerful form of regression that can account for both fixed and random effects, is robust against several assumption violations, and is flexible in the kind of variables it can take (Cunnings, 2012).

Two features of mixed models are useful here. First, a mixed model can model out the effect of author on the use of linguistic features. That is, by specifying author as a random effect, the model can account for differences in the dependent variable due to authors' writing style, producing more accurate estimates for the fixed effects of interest. Second, a mixed model can include interaction terms, or the effect of one variable on the dependent variable depending on the value of another variable. For example, the SNA corpus may show significantly greater use of modal verbs in health texts relative to the RA corpus, an effect not captured by comparing the effect of register alone.

LME models were run using the lmer() function in the *lme4* package (Bates et al., 2015) in R. Effect sizes are reported using conditional R^2 values, which express what percentage of variance in the outcome variable explained by the independent variables (Nakagawa et al., 2017).

3.1.5.5 Grouping texts together with cluster analysis

In chapters 5 and 6, many linguistic features are examined. For example, chapter 5 examines 14 different verb patterns. While all features are first examined separately, I also adopt a method called *k*-means cluster analysis to group texts based on linguistic features. *K*-means cluster analysis assigns *k*-random points in a two-dimensional space, identifies the data that are closest to it based on some criterion, re-calculates the middle of the group of data points, re-adjusts which data are to be clustered with the new centers, and iterates on this process until no more changes in the clusters occur.

K-means is an intuitive and interpretable clustering algorithm widely used by researchers (Moisl, 2015). However, two weaknesses include the fact that the researcher must specify how many clusters to search for, as well as where to position the initial centers in the space. Computational techniques help to minimize these issues. Specifically, the *hclust()* function in R can be used to produce a hierarchical dendrogram visualizing the optimal number of clusters in the data. That number can then be used as the *k* in a *k*-means analysis. In addition, the *kmeans()* function offers a parameter, *nstart = n*, which can be used to run *k*-means analysis *n* times, selecting the solution that best minimizes error (i.e., total sum of squares).

Finally, there are other issues and assumptions worth noting. First, *k*-means cluster analysis is sensitive to outliers. Thus, the dendrograms produced in chapters 5 and 6 were used to detect any individual texts that had undue influence on the clusters. Second, it has been found that the number of observations in a dataset should be 70 times the number of variables included in the cluster analysis (Dolnicar et al., 2014). Thus, with regard to the verb pattern cluster analysis, only the 10 most frequent verb patterns are included in (70 * 10 variables = 700 observations needed). Finally, the *vif()* function in the *car* package (Fox & Weisberg, 2019) was used to test the assumptions of multicollinearity and singularity and were found to be not violated.

3.2 Semi-structured interviews

As noted earlier, this dissertation adopts an explanatory sequential design, where interviews with SNA writers are used to inform the interpretation of the corpus study results. Specifically, I adopt discourse-based interviews (Conrad, 2014) in which informants are used to help elucidate the findings of a corpus linguistic study. Below, I describe the interview design and procedure.

3.2.1 Interview type and interview guide

I conducted semi-structured interviews with several writers of SNAs. Semi-structured interviews tend to be useful for qualitative analysis because of their flexibility (Braun & Clark, 2013). This kind of interview also aligns with my ontological perspective on this research, which sees meaning making as socially constructed rather than "out there" to be found (Glesne, 2014).

I designed an interview guide using principles and considerations described in Braun and Clarke (2013, p. 81-85), including utilizing open-ended questions, designing opening and closing questions, scrutinizing and revising questions, and considering question sequence, among other things. The example guide provided in Clarke (2006) served as an exemplar to build my own. The resulting guide includes about 10 open-ended questions about the participant's experiences writing science news, as well as questions focusing on the results of the corpus analysis (Appendix D).

3.2.2 Participant Sampling

To identify participants, I adopted a typical case sampling approach, wherein participants are recruited on the basis that they represent the average participant of the larger group (Patton, 2015, p. 405-406). While identifying the average participant of a group may seem problematic, an average or typical participant in this dissertation was operationalized by recruiting one writer from each of the eight SNA sources. A list of the writers who contributed the most texts to each of the eight sub-corpora was created. I contacted writers from each list one at a time until one writer from each sub-corpus agreed to take part. However, no writers from the outlet NewAtlas agreed to be interviewed and secondary interviewees from other outlets were also difficult to recruit, so seven, rather than eight, participants from seven different outlets were recruited.

Each interview lasted between 40 and 55 minutes, took place over video call, and was audio recorded. Audio recordings were then transcribed verbatim. The letter of approval by the Institutional Review Board at Georgia State University for this interview study can be found in Appendix E.

4 SITUATIONAL CHARACTERISTICS ANALYSIS

In this section, I answer the first research question stated in the introduction (Chapter 1), namely, what are the situational characteristics of the SNA corpus and how do they compare and contrast with those of the RA corpus? Describing the context around which language is produced helps link linguistic choices with their functions in discourse. In the register studies tradition, such a description is called a situational characteristics analysis (Biber & Conrad, 2019). In section 4.1, I describe a situational characteristics framework for the texts in the SNA corpus, followed by a framework for the texts in the RA corpus in section 4.2.

4.1 A framework for the situational characteristics of the SNA corpus

Biber and Conrad (2019) surveyed previous frameworks for analyzing registers and identified seven features that are generally useful for many situational characteristics analyses, including participants, relations among participants, channel, processing circumstances, setting, communicative purposes, and topic. Each feature has two or more sub-characteristics to flesh out the description. For example, within participant, a researcher would differentiate between who the addressor is, who the addressee is, whether there are additional on-lookers, and so on.

This general framework is useful for distinguishing high-level differences between registers, but describing more specific differences, especially between registers which are similar, necessitates alterations to this framework. When developing the framework for the texts in the SNA corpus, I kept in mind Conrad’s (1996) suggestions that a framework should be independent from linguistic features, comprehensive, applicable to many texts, and applicable to comparisons between corpora at a variety of levels. The resulting framework is shown in Table 4.1.

Table 4.1 A situational characteristics framework for the SNA corpus

Characteristic	Options
1. Participants	
A. Writers	1 (single authored) 1+ (multiple authors)
B. Readers	Locale of readership Frequency of website visits (Named) intended audience
2. Textual layout & organization	
A. Text length	Mean length Standard deviation of length
B. Headings	Some headings None
3. Digital affordances	
A. Visual elements	One still image Still images Multimedia (videos, GIFs) None
B. Number of hyperlinks	Mean per text Standard deviation per text
C. Source of hyperlinks	Number of hyperlinks linking to professional science sources Number of hyperlinks linking to other sources
D. Reader interactivity	Allows for readers to comment on article and/or share article on social media None
4. Setting	
A. Outlet legacy	0-10 years in press 10-50 years in press 50 + years in press
B. Modes of publication	Published online only Published both in print and online

C. # of employees	2-10 employees 11-50 employees
D. Nature of outlet	Articles focused on academic research Articles focused both on academic research and other entertainment
5. Communicative purpose	
A. General purpose	e.g., narrate/report, describe, inform/explain/interpret, persuade, how-to/procedural, entertain, edify, reveal self
B. Specific purpose	e.g., summarize information from numerous sources, present new research findings, teach moral through personal story
6. Subject/topic	
A. Nature of news	Articles focus on a specific academic field(s) Articles focus on several academic fields

This framework has six main characteristics and several sub-characteristics. Below, I describe each in more detail.

4.1.1 Participants

The first category is a description of the text producers and audience. It involves two sub-characteristics, including the number of authors per text and the intended audience. Ideally, more information about the authors’ backgrounds would be included, but the large number of authors and minimal information available online makes inclusion of this information difficult. Finally, this category also includes information about who the intended audience is, including where the audience is located, how frequently they read the news outlet, and how the outlet describes their audience on their website.

4.1.2 Textual Layout and Organization

The second feature includes six sub-characteristics describing layout and organization and includes two features, namely the average length of texts and the presence or absence of headers or sub-titles within the body of texts. The use of headings is an important indicator of functional differences within parts of research articles, so their presence of absence within SNAs is also a relevant feature to explore.

4.1.3 Digital affordances

Digital affordances refer to features that distinguish a print text from a digital one. As discussed in section 2.1, digital genres offer certain affordances by way of their being on the Internet. For this analysis, three features were included for consideration, namely the presence and type of visuals included in texts, the use of hyperlinks to link to external sources, and the ability for readers to interact with the content by leaving comments or sharing the article on social media via a ‘share’ icon.

4.1.4 Setting

The setting characteristic includes the number of years the news organization has been in business, the mode in which the outlet publishes its articles, the number of employees, and the subject area that the outlet generally reports on. Information for these characteristics were gathered from online websites, such as LinkedIn, the outlet’s website, and third-party websites that report statistics about online traffic. These characteristics were useful for comparing the different SNA outlets.

4.1.5 Communicative purpose

Communicative purpose describes the general and specific purpose of the register. Often, the general purpose aligns with frequently used names for genres, such as narrative, persuasion, or how-to. A specific purpose may comprise one part of a text, such as the specific purpose of the methods section in a research article or may include fine-grained purposes that add detail to the general purpose. In short, the communicative purpose answers *why* a register communicates its message.

4.1.6 Subject/Topic

Subject/topic includes one sub-characteristic, namely whether the news outlet specializes in writing about a single domain, such as psychology, or writes about multiple domains, such as health, astronomy, and psychology. This characteristic was important because initial review of outlets found that they often differ with regard to their specificity of subject matter.

4.1.7 The Situational Characteristics of the SNA corpus

Table 4.2 presents the situational characteristics analysis of the SNA corpus, grouped by topical category, using the framework described in the previous sections. Following the table, I briefly describe the texts in the SNA corpus by their situational characteristics.

Table 4.2 The situational characteristics of the SNA corpus

Characteristic	Topical category			
	Space	Environment	Health	Psychology
1. Participants				

A. Writers				
Single author	100	100	100	100
Multiple authors	0	0	0	0
B. Readers				
% U.S.-based	46-53%	40-51%	48-53%	48%
Monthly web visits	1-10 mil	1-19 mil	2-4 mil	4 mil
(Named) intended audience	Enthusiasts	General public	General public	General public +, academics
2. Textual layout & organization				
A. Text length				
M length	860.23	692.69	566.95	694.57
SD length	240.15	135.36	243.85	257.06
B. Headings				
Some present	77	15	3	0
None	23	85	97	100
3. Digital affordances				
A. Visual elements				
1 still image	16	54	73	100
1+ still images	50	40	20	0
Multimedia	34	6	7	0
None	0	0	0	0
B. Number of hyperlinks				
M hyperlinks	7.39	10.92	3.39	2.52
SD hyperlinks	4.13	4.14	1.92	1.09
C. Sources of hyperlink				
% research-related	70.24%	53.74%	46.15%	100%
% other	29.76%	46.26%	53.85%	0
D. Reader interactivity				
Present	0	1	2	1
None	2	1	0	1
4. Setting				
A. Outlet legacy				
0-10 years	1	0	0	1
10-50 years	1	2	1	1
50 + years	0	0	1	0

B. Modes of publication				
Online only	1	2	1	2
Print and online	1	0	1	0
C. # of employees				
2-10	0	2	0	1
11-50	2	0	2	1
D. Nature of outlet				
Research focused	0	2	1	2
Research + other entertainment	2	0	1	0
5. Communicative purpose				
A. General purpose	To inform and entertain			
B. Specific purpose	Summarize and explain research findings from a recently published research article			
6. Subject/topic				
A. Nature of news				
Specific field	1	0	0	1
Broad fields	1	2	2	1

4.1.7.1 Participants

As noted in section 3.1.1, only SNAs with single authors were considered for collection. Naturally, Table 4.2 reflects this fact. Additionally, it is worth noting that there is no evidence that these texts were written using generative Artificial Intelligence. Few public tools (e.g., Open AI's *ChatGPT*) were available when the texts were published, and no disclaimers identifying that the texts were generated via AI were found in the corpus. Regarding intended audience, outlets' websites and informants noted having either a general or an enthusiast readership. In the case of one source, the intended audience included teachers, researchers, and students, for whom the SNAs served to keep readers up to date in the fast-moving world of academia.

Online data suggested that about half of all online readers for the eight news outlets were U.S.-based (as of summer 2022). However, the number of monthly visitors varied between outlets. For example, ZME Science boasted about 1 million monthly visitors, PsyPost boasted about 4 million, and Science Alert about 19 million. However, different sources reported different figures, so these numbers should be taken with caution.

4.1.7.2 Textual Organization

The average SNA was 703.61 words with a standard deviation of 246.75 words. The longest text was a space article 2,486 words in length, while the shortest was a health article 279 words in length. Health articles were the shortest on average and space articles were the longest. While most space articles used headings, psychology, environment, or health articles rarely did.

4.1.7.3 Digital affordances

Most SNAs displayed one or more still images, while a smaller number utilized multimedia in the form of videos or GIFs. When a single photo was included, it was most often displayed just above the article's title. SNAs generally hyperlinked 3-11 sources in their texts, often to both academic and non-academic sources. Non-academic sources were often other articles published by the same news outlet, which served to explain important concepts for the given article or review other relevant research. Finally, outlets ranged in their interactivity with readers. Half of the outlets either provided comments sections for readers or quick ways to post the article on social media, while the other half provided neither.

4.1.7.4 Setting

At the time of investigation (summer 2022), two outlets had been established for less than 10 years, five had been established for between 10 and 50 years, and one had been established for more than 50 years. Six sources published online-only content, while another two published both online and in print. Online data suggested that three outlets had fewer than 10 employees and the remainder employed between 10 and 50. Again, these numbers should be taken with caution.

Finally, three outlets covered both academia and casual entertainment. For example, one outlet published articles on the topics of innovation, mind and body, gaming, and culture, while several others reported only on academic research, such as humans, Earth, space, and physics.

4.1.7.5 Communicative purpose

Several informants described the purpose of SNAs being to inform and entertain. For example, the informant quoted below describes how SNAs must balance entertainment value with accuracy, while RAs mainly focus on detail and precision:

> *The goal of news articles is that they are entertaining, and that they're truthful and that they're precise while not sacrificing their entertainment value […] while in science writing, your goal is that your text is as precise as possible, and as detailed as possible, because the idea of scientific text is that you […] include in your text enough data and enough detail so that someone can replicate your study* [Informant #1]

Thus, SNA writers must strike a balance between attracting readers and remaining accurate to the source material. In short, the general purpose is to entertain readers while also keeping them informed on the world of scientific research. While some SNAs are likely multi-functional, the most common specific purpose of this register can be described as summarizing recently published academic research.

4.1.7.6 Subject/Topic

Most outlets included in this study wrote about multiple academic domains, from anthropology to medicine to mental health, while two specialized in a single domain, namely space and psychology. Sometimes, outlets tagged articles with more specific topics. For example, the outlet specializing in psychology also tagged texts with topics such as psychedelic drugs, conspiracy theories, relationships, sexual health, and Donald Trump. Thus, even outlets which focused on a particular domain wrote on a variety of topics therein.

4.1.7.7 Summarizing trends in the situational characteristics of the SNA corpus

The above situational characteristics analysis of the SNAs suggests a good deal of internal consistency within the corpus. All articles were single authored, written for a largely U.S.-based general audience, spanned about 700 words, included some visual elements, hyperlinked to external sources, and wrote about recently published research. However, one source noted a more academic intended audience, thus challenging the standard ‘general audience’ motif of popular science.

Notable differences were found in the settings characteristic. Here, the number of monthly website visits varied greatly between outlets, which also varied in the number of years that they had been operating at the time of investigation. Another notable difference was found in the visual layout of SNAs. While some articles were accompanied by several pictures, figures and graphs, and even video clips, others were largely unadorned, with just a single static image above the title and no other sub-headings. Similarly, many outlets did not utilize the affordance of reader interactivity provided by Web 2.0 technology, suggesting that these websites were a place to consume articles but not discuss them.

4.2 The Situational Characteristics of the RA Corpus

In this study, I rely on the framework developed by Gray (2015) to analyze the situational characteristics analysis of the RA corpus. Gray’s framework offers eight characteristics with up to five sub-characteristics in each. In the framework shown below (Table 4.3), many features were adopted from Gray, while others were slightly modified, removed, or added to better match the needs of this study. For example, due to the large number of unique journals in the RA corpus, it was not feasible to carry out a survey of all journals. Instead, the 10 most frequently cited journals from each sub-corpus were examined for whether they were a generalist or specialized journal. Similarly, for subject/topic and object of study, a random sample of 30 articles from each sub-corpus was collected and surveyed.

Table 4.3 The Situational Characteristics of the RA Corpus

	Space	Environment	Health	Psychology
1. Participants				
A. # of authors				
1 author	2	1	1	0
2-4 authors	35	42	11	61
5+ authors	63	57	88	39
2. Textual layout & organization				
A. Text length				
M length	4859.19	4725.72	4587.14	5970.85
SD length	2826.46	1868.36	1969.93	2654.5
B. Headings				
Un-numbered	42	71	83	63
Numbered	44	24	13	37
None	14	5	4	1
C. Visual elements				
Visuals (-equations)	46	64	82	88
Visuals (+equations)	53	36	18	10
None	1	0	0	2
D. Sections/organization				
IMRD	27	36	24	96
Varied IMRD	13	26	59	1
Other	46	33	13	2
None	14	5	4	1
3. Setting				

A. Number of journals	33	81	47	79
B. Nature of journal				
Generalist	7	7	7	6
Specialized	2	3	3	4
Other	1	0	0	0
4. Subject/topic				
A. General topics	Physical, extraterrestrial world	Living organisms and their relationship to Earth	Physical workings of the human brain, human diseases and health	Understanding human behavior
5. Communicative purpose				
A. General purpose				
To present analysis of observed data	93	97	100	95
Other	7	3	0	5
6. Nature of data/evidence				
A. Presence of observed data				
Yes				
No	93	97	100	95
	7	3	0	5
B. Use of numerical evidence				
Yes	98	99	99	95
No	2	1	1	5
C. Object of study	Laboratory data, measures of natural phenomena, numerical models & formulations	Modeling and simulations, modeling, survey data, historical	Government statistics, laboratory data, simulations, modeling, numerical	Government statistics, surveys, experimental conditions, video recordings, logical

		artifacts, measures of natural phenomena	formulations, measures of natural phenomena	progression, language production
7. Methodology				
A. General method				
Empirical	93	96	98	96
Non-empirical	7	4	2	4
B. Statistical techniques				
Descriptive stats	0	22	12	1
Statistical differences	0	13	25	13
Other advanced stats	98	61	59	81
None	2	4	4	5
8. Explicitness of research design				
A. Explicitness of RQs				
Direct statement	0	3	0	25
Minimal/none	100	97	100	75
B. Explicitness of Citations				
Within text	43	27	8	87
In notes	56	73	92	13
None	1	0	0	0
C. Explicitness of Evidence				
Extensive				
Mention / none	95	100	96	99
	5	0	4	1

D. Explicitness of Procedures				
Extensive				
Sep. / min. / none	83	87	87	99
	17	13	13	1

4.2.1.1 Trends in the situational characteristics of the RA corpus

The RA corpus represents the RAs that the SNAs reported on. Despite this unusual selection criterion, the RAs display expected similarities and differences. Among the similarities, most RAs were authored by more than one writer, and the mean lengths of texts were similar across sub-corpora, save for psychology RAs, which were ~1000 words longer than the other texts. Most RAs use headings to separate sections and show tables and figures. Most RAs were empirical in nature, reporting on observed data using numerical reasoning in the form of basic or advanced stats, for the purpose of contributing new knowledge to an academic field.

Regarding differences, each sub-corpus focused on different topics. Space RAs examine the physical environment beyond Earth, while environment RAs examine the physical environment of Earth and humanity's relationship to it. Health RAs examine human health and the physical workings of the brain, and psychology RAs examine human behavior. Naturally, these sub-corpora also vary in their objects of study, with some RAs examining laboratory data and others surveys data. Finally, the sub-corpora differ with respect to sectional organization. While psychology RAs largely follow the prototypical Introduction-Methods-Results-Discussion (IMRD) organization, others varied in their organization, especially by placing the methods section at the end of the article or removing it altogether.

4.2.1.2 Comparing the situational characteristics of the RA and SNA registers

The characteristics of the RAs and SNAs show important differences. One difference is that RAs are much longer, with well-defined sections fulfilling the needs of specific aspects of research writing. SNAs are shorter and easily digestible by way of their having few visual interruptions within their short bodies of text. Additionally, the RAs were written by researcher teams who specialize in the subject area, while SNAs were written by a single journalist or freelance writer with varying degrees of subject-area experience. Finally, the purpose of most RAs is to present the results of empirical research and thereby contribute to a body of literature, while SNAs summarize those findings after publication and connect them to the readers' lives. In the next three chapters, I will refer back to these characteristics to make functional links between textual findings and their uses in SNAs and RAs.

5 VERB PATTERN ANALYSIS

In this chapter, I describe findings from the analysis of verb patterns. Specifically, I seek to answer which verb patterns are most descriptive of each register, how those patterns suggest similarities or differences between the registers, and what functions they may fulfill in discourse, using corpus data and informant data to inform analysis. Table 3.5 in section 3.1.4.1 displayed 30 verb patterns that were considered for analysis. To more easily compare and visualize the data, these 30 patterns were condensed into 14 more inclusive patterns, shown below:

1. Intransitive patterns (IV+0)
2. Reporting clause patterns (FCl+S|TV)
3. Existential *there* patterns (Ex*There*+*BE*)
4. Interrogative clause patterns (Question)
5. Copular patterns with phrasal complements (CV+PhC)
6. Copular patterns with finite complement clauses (CV+FCl)
7. Copular patterns with non-finite complement clauses (CV+NfCl)
8. Passive patterns without complement clauses (PassTV)
9. Passive patterns with finite complement clauses (PassTV+FCl)
10. Passive patterns with non-finite complement clauses (PassTV+NfCl)
11. Transitive patterns with phrasal objects (TV+PhO)
12. Transitive patterns with finite complement clauses (TV+FCl)
13. Transitive patterns with non-finite complement clauses (TV+NfCl)
14. Transitive patterns with phrasal objects and non-finite complement clauses (TV+PhO+NfCl)

This chapter's findings will be organized as follows. In section 5.1, I report on the distributions of the 14 patterns across the registers. In 5.2, I explore the use of patterns with transitive verbs, including those patterns with phrasal objects and complement clauses, before turning to patterns with copular verbs in 5.3. In section 5.4, I describe results of a cluster analysis to explore how these patterns work together to form groups of like-texts. Finally, 5.5 explores the use of optional adverbials across the verb patterns.

5.1 The distribution of verb patterns across the registers

The distributions of the 14 verb patterns across the registers are shown in two bar graphs displayed in Figure 5.1.

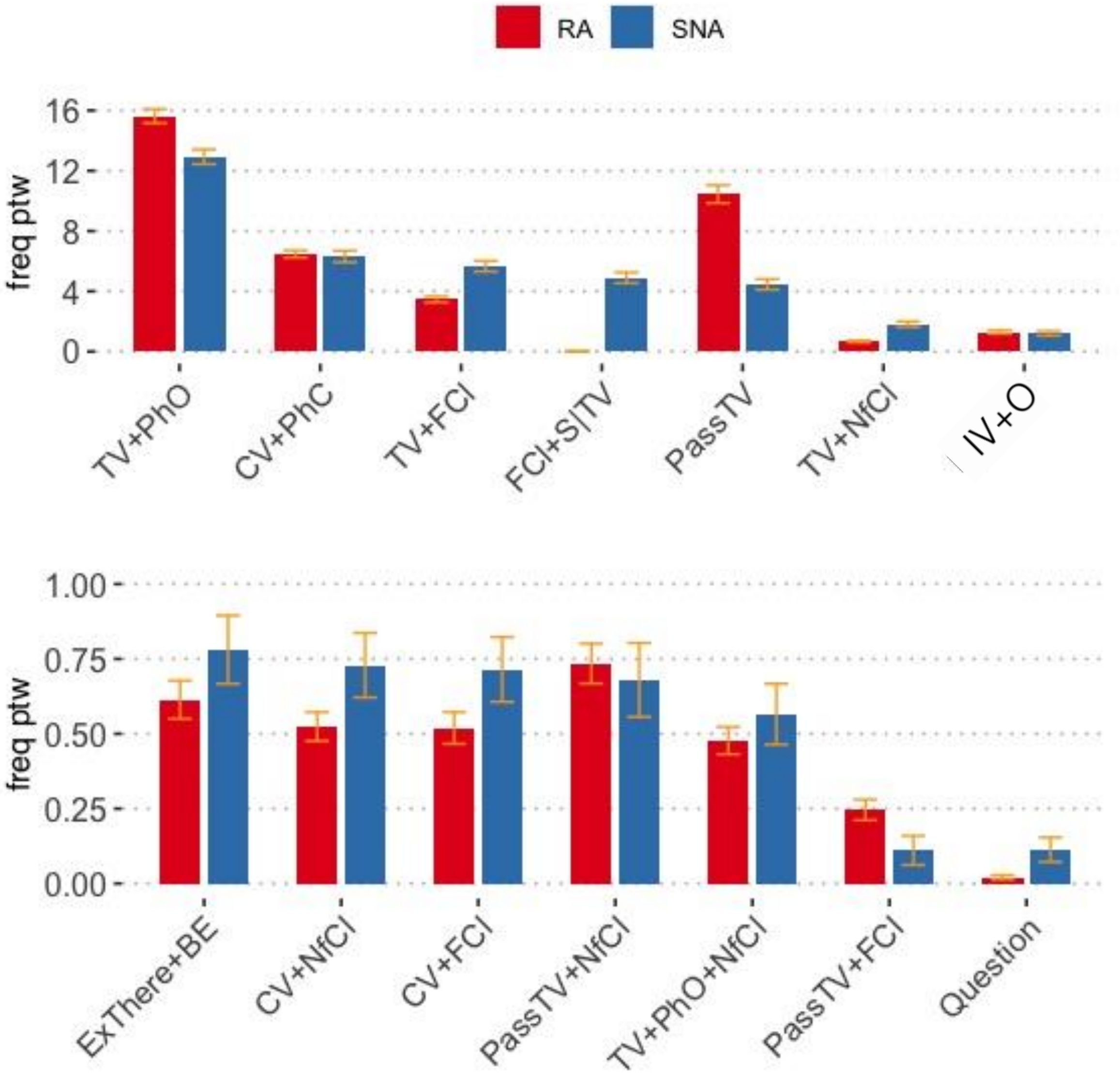


Figure 5.1 (Top) A bar graph illustrating the 7 most frequent verb patterns across the registers. (Bottom) A bar graph illustrating the next 7 most frequent verb patterns across the registers

The 95% confidence intervals (CIs) shown in Figure 5.1 show significant differences in the rates of transitive verbs with phrasal objects (TV+PhO) and two passive voice transitive patterns (PassTV, PassTV+FCl), such that RAs used more of these patterns. By contrast, SNAs appeared to use significantly more transitive verbs with complement clauses (TV+FCl, FCl+S|TV, TV+NfCl) and copular verbs with complement clauses (CV+NFCl, CV+FCl), as well as interrogative clauses (Question). The most meaningful differences were among the use of the PassTV (*d* = 1.21), TV+NfCl (*d* = 0.84), and TV+FCl (*d* = 0.76) patterns, which showed large effect sizes. TV+PhO patterns showed a moderate effect size (*d* = 0.55), while PassTV+FCl (*d* = 0.32), CV+NfCl (*d* = 0.24), and CV+FCl (*d* = 0.23) patterns showed smaller effects.

Another way of visualizing these patterns is by their proportional use. The pie charts in Figure 5.2 visualize the proportional use of verb patterns in a prototypical RA and prototypical SNA.

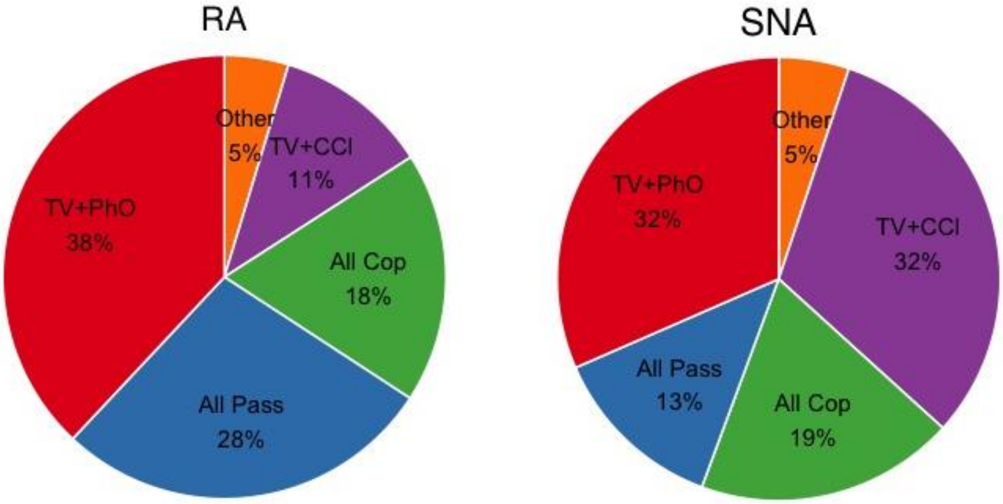


Figure 5.2 Two pie charts illustrating the proportional use of verb patterns in each register

The pie charts displayed in Figure 5.2 show that, when reading a prototypical SNA, a reader will likely come across roughly equal parts transitive verbs with phrasal objects and transitive verbs with complement clauses. The remaining one-third of verb patterns will include mostly copular and passive voice verbs. By contrast, the prototypical RA will likely show more transitive verbs with phrasal objects, including those in the active (TV+PhO) and passive (All Pass) voice, largely at the expensive of transitive verbs with complement clauses.

To better understand how these patterns were used in discourse, I next turn to exploring a number of verb patterns in more detail, specifically those that were frequent and those that differed significantly across the registers. In these analyses, I discuss the semantic domains of the patterns' clausal elements, specifically the animicity of their subjects (e.g., *scientist* vs. *model*) and the semantic domains of their verbs (e.g., mental verbs, verbs of communication). I adopt Biber et al.'s (1999) verb domains to analyze those in this dissertation.

In section 5.2, I focus on transitive patterns, looking first at those with phrasal objects and then those with complement clauses. Then, I turn to copular patterns in section 5.3.

5.2 Transitive verb patterns

Patterns with transitive verbs were more frequent than any other pattern. However, the registers differed with regard to which transitive patterns they used most. In particular, the SNA register showed a strong preference for transitive patterns with complement clauses, while the RA register showed a strong preference for transitive patterns with phrasal objects, as well as their passive forms. In section 5.2.1, I focus on active voice TV+PhO patterns, before turning to passive voice PassTV patterns in 5.2.2.

5.2.1 Transitive patterns with phrasal objects

Transitive patterns with phrasal objects came in four varieties, namely those with a single phrasal object headed by a common noun, pronoun, or proper noun, and those with two phrasal objects, typically referred to as ditransitive. Their distribution across the registers is shown in Figure 5.3.

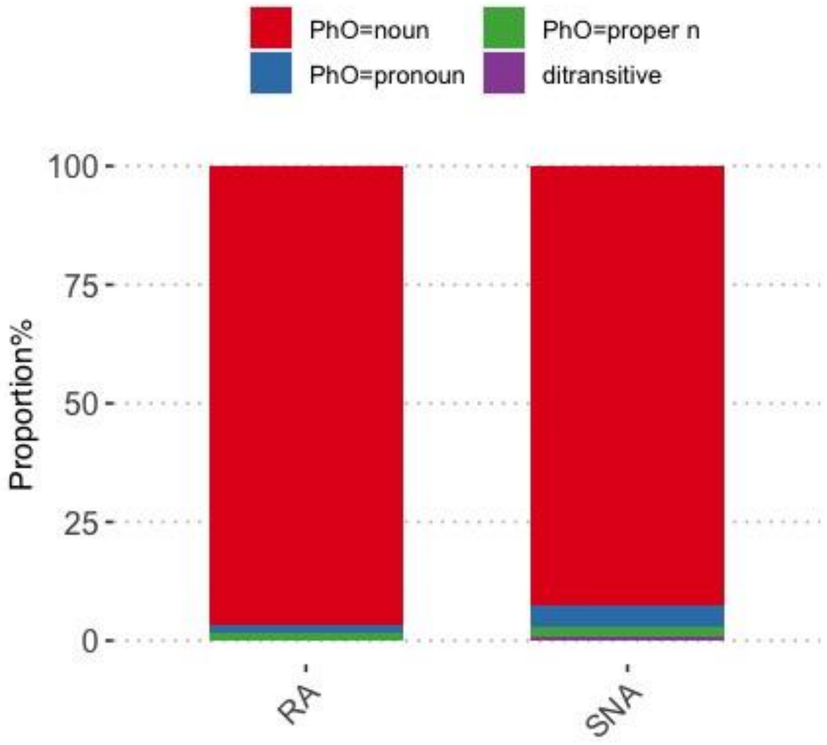


Figure 5.3 A bar graph illustrating the proportional use of four varieties of TV+PhO patterns across the registers

Figure 5.3 shows that both registers relied heavily on phrasal objects headed by common nouns, while only those headed by pronouns were somewhat noticeable in the SNA corpus (~5% of all TV+PhO patterns). The most frequent subjects, verbs, and direct objects of TV+PhO patterns suggest that these sentences were often used to report technical details of research, especially in RAs. SNAs often used this pattern to describe research results or provide topical background information to contextualize a study. First, consider the most frequent direct objects in the TV+PhO pattern (Table 5.1).

Table 5.1 The ten most frequent direct objects of TV+PhO patterns across the registers

RA		SNA	
direct object	freq ptw (raw)	direct object	freq ptw (raw)
effect	0.26 (474)	**it**	0.17 (49)
model	0.18 (337)	data	0.16 (45)
data	0.14 (265)	effect	0.15 (42)
result	0.13 (247)	evidence	0.12 (33)
analysis	0.13 (242)	**study**	0.09 (26)
evidence	0.13 (232)	**they**	0.09 (24)
difference	0.12 (224)	**way**	0.09 (24)
value	0.11 (194)	difference	0.08 (22)
association	0.1 (188)	**sample**	0.08 (22)
level	0.1 (184)	**question**	0.07 (21)

Note: **bolded** words are those that are unique to their respective column

The most frequent direct objects, especially those that only appear in the RA column, suggest a relationship with statistical descriptions. Often, patterns with these objects were found in the methods sections of RAs, a conclusion also supported by the most frequent subjects of TV+PhO patterns (Table 5.2).

Table 5.2 The ten most frequent subjects of TV+PhO patterns across the registers

RA		SNA	
subject	freq ptw (raw)	subject	freq ptw (raw)
we	4.1 (7522)	**researcher**	0.89 (250)
study	1.26 (726)	they	0.81 (229)
this	0.29 (533)	**team**	0.59 (166)
participant	0.25 (456)	it	0.56 (159)
they	0.17 (317)	we	0.55 (156)
analysis	0.17 (313)	study	0.47 (133)
result	0.17 (312)	**scientist**	0.3 (85)
model	0.17 (304)	this	0.21 (60)
it	0.16 (301)	research	0.15 (43)
research	0.13 (241)	participant	0.14 (40)

Note: **bolded** words are those that are unique to their respective column

Table 5.2 shows that, while there are many ways of referring to humans in the SNA corpus (e.g., *researcher, team, scientist*), the RA corpus largely relied on *we* for this function, where it was used more than three times as often as any other subject in that register. Thus, a typical TV+PhO pattern in the RA corpus consists of the subject *we* alongside an object referring to a methodological step. Excerpts (5.1) and (5.2) below exemplify these characteristics (subjects **bolded**, direct objects underlined):

(5.1) Using LISREL, **we** carried out Path analyses to test the model. [Psychology RA #154]

(5.2) Finally, **we** ran a model controlling for sexual frequency. [Psychology RA #240]

The fewer instances of patterns like (5.1) and (5.2) in the SNA corpus is likely due in part to the lesser focus on research methodology in SNAs in general. Indeed, informants noted that methodology was often inconsequential to writing a good SNA. For example, one informant connected methodological information to uninteresting, technical writing:

> *In science news articles, methods are extremely simplified. And because it is my expectation that even if I tried to present the methods truthfully, the audience would not understand it, so it's a waste of time and it would make the article boring and hard to read. So, basically, you present some key elements of the methods but do not go into the fine details.* [Informant #1]

Looking at TV+PhO patterns in the SNA corpus, rather than focusing on what researchers did to what in the methodology, writers instead focused on study findings, using animate or inanimate nouns as their subjects. Excerpts (5.3) and (5.4) illustrate these characteristics (subjects **bolded**, direct objects underlined):

(5.3) **New research** provides evidence that democratically elected leaders tend to have more attractive and warmer faces than authoritarian leaders. [Psychology SNA #300]

(5.4) But **the team** found no effects of the drug on nerve cells' dendritic spines - tiny signal-receiving blebs that help make new neural connections. [Health SNA #145]

These excerpts highlight the outcomes of scientific research, such as evidence for a theory or a measurable impact of a drug. They also show that SNA writers can pack a lot of information into the phrasal objects. That is, while the direct object is a noun phrase, the noun head is often modified or complemented by stacked phrases or clauses, making the object informationally dense. For example, the head of the object in (5.3), *evidence*, is complemented by a finite *that* clause, allowing the writer to unpack what the evidence is with a grammatically explicit verb phrase and other clause elements.

In addition to study findings, TV+PhO patterns in the SNA corpus also provided topical background information to contextualize the study under review. Excerpts (5.5) and (5.6) below illustrate this use (subjects **bolded**, direct objects underlined).

(5.5) **The immune system** normally does a pretty good job of fighting off disease, but cancer is known to use all kinds of sneaky tricks to evade detection. [Health SNA #78]

(5.6) Some of [our plastic] ends up in landfills, a lot of it ends up in the oceans. **We**'ve seen it in every part of the oceanic food chain, and **we**'ve seen it in every corner of the ocean. [Environment SNA #363]

Both excerpts come from the start of their respective SNA, setting the context for how a recent study changed or updated the current understanding of, for example, how the immune system fights cancer or how much plastic is polluting the ocean.

5.2.2 Transitive patterns with passive voice verbs

Transitive patterns with passive voice verbs came in three varieties, namely those without a post-verbal complement, those with a *to* clause complement, and those with a *that* clause complement. Their distribution across the registers is shown in Figure 5.4.

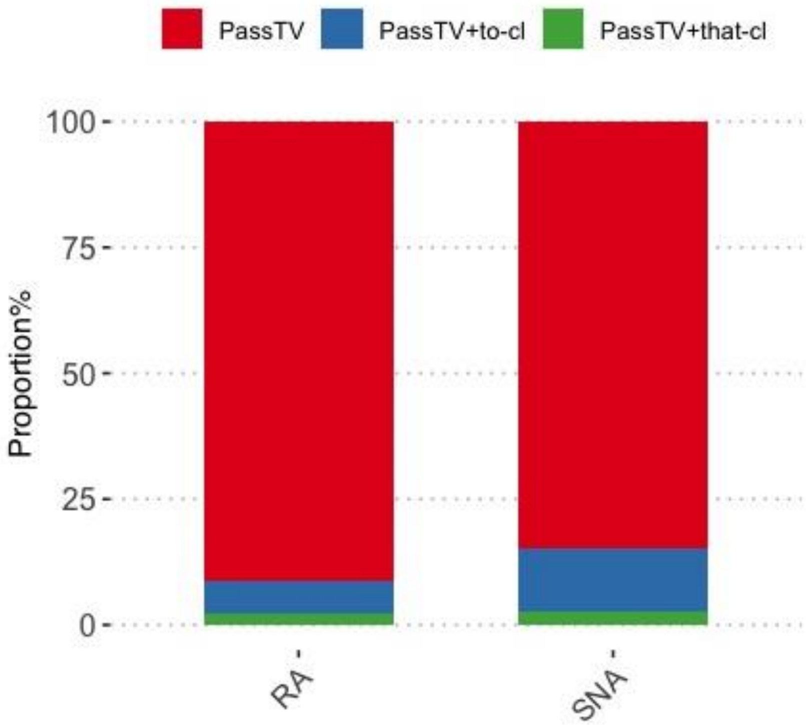


Figure 5.4 A bar plot illustrating the proportional use of three varieties of passive transitive patterns across the registers

Figure 5.4 shows that PassTV patterns were heavily preferred relative to those with a complement clause, though the SNA corpus relied on PassTV+*to*-cl patterns to moderate degree (~12% of all passive voice transitive patterns). In either register, mental verbs like *find*, *expect,* and *think* were among the most common verbs in passive patterns with a complement clause. However, the verb *note*, serving as a polite directive to the reader, stood out as being frequent in RAs but absent from SNAs. Excerpts (5.7) and (5.8) demonstrate the use of mental verbs and *note* in passive patterns with complement clauses (passive verbs **bolded**, complement clauses <u>underlined</u>):

(5.7) Despite spreading questionable information, fake news articles **were found** <u>to be more likely to go viral which further increases their reach in society</u>. [Psychology SNA #210]

(5.8) It **should be noted** that the purpose of modeling here is to only make an estimate of the long-term trend. [Space RA #69]

Excerpt (5.7) includes a *to* complement clause controlled by the mental verb *find*, a pattern relied upon more by SNA writers. By contrast, sentences with a *that* complement clause controlled by *note* were largely found in RAs, where they functioned to politely instruct readers to consider some information in a particular way (e.g., *note that the purpose of* X *is* Y).

However, most patterns with passive transitive verbs did not include a complement clause. Instead, most of these patterns required a subject, corresponding to the direct object in the active voice, and the verb. Each register was fairly idiosyncratic in its reliance on certain passive verbs, as Table 5.3 displays.

Table 5.3 The ten most frequent passive voice transitive verbs across the registers

RA		SNA	
verb	freq ptw (raw)	verb	freq ptw (raw)
use	0.57 (1051)	**publish**	0.63 (177)
show	0.26 (474)	**author**	0.29 (81)
perform	0.26 (472)	find	0.23 (64)
find	0.25 (457)	use	0.12 (35)
observe	0.24 (438)	**ask**	0.09 (25)
associate	0.2 (374)	**know**	0.09 (25)
calculate	0.17 (307)	associate	0.08 (23)
measure	0.15 (277)	**think**	0.08 (22)
conduct	0.15 (276)	**call**	0.07 (20)

consider	0.13 (238)	**link**	0.06 (18)

Note: **bolded** words are those that are unique to their respective column

Table 5.3 shows that seven of the top ten verbs from either register were unique to their respective column. In other words, many of the most common verbs in one register were not also the most common verbs in the other register. Several verbs in the RA column highlight the physical activities associated with conducting research, such as *perform, calculate,* and *measure*, the implied agents of which are the researchers themselves, such as in (5.9) (passive verb **bolded**):

(5.9) The background **was measured** in 2.1-px radius apertures randomly placed at the same stellocentric distance as the target source. [Space RA #137]

Much like excerpts shown in section 5.2.1, excerpt (5.9) demonstrates the presence of methodological description in the RA corpus and its corresponding weaker prominence in the SNA corpus.

By contrast, SNA verbs in Table 5.3 are quite diverse. Although *ask* explains what participants were requested to do in experiments, the two most frequent verbs, *publish* and *author*, are specific to the needs of SNAs. They signal in which academic journal an article was published (*was published in*) and who wrote it (*was authored by*), often appearing near the top of articles. Their position in articles reflects the brief and compact nature of SNA writing, which functions to identify recent, newsworthy research, provide only the relevant details, and then move on to the next piece of news. One informant put it this way:

> *We're writing things that are so short, there's very little space to get deep into the background stuff. Particularly because the background stuff, if it's there, it is towards the*

> *top of the story, and we always want to get people through that as quickly as possible, so we can get into the meat of the actual study that we're covering.* [Informant #5]

Informing readers about what the study is called, where it was published, and who wrote it is a more-or-less obligatory feature of single-study SNAs because readers expect the article to focus largely on the relevant details of the new research. Longer digressions and deeper context are left for other genres, registers, and contexts of science communication.

Additionally, a pair of mental verbs (*know, think*) and a verb of communication (*call*) were found in the SNA column but not the RA column. The contexts around these verbs suggested that they were often used to identify a general consensus about a subject matter. Consider excerpts (5.10) and (5.11):

(5.10) This condition **is known as** orthostatic hypotension, which is a temporary drop in blood pressure. [Health SNA #116]

(5.11) Gossip **is** generally **thought** to be a bad thing, especially in the workplace, but a new study has found that positive gossip among teams can have some benefits […] [Psychology RA #123]

Sentences like these demonstrate audience awareness on the part of SNA writers. By defining terms and defining the current state of knowledge, the writer constructs a common ground between themselves and the reader which they can continue to develop in the article. Additionally, placing these verbs in the passive voice helps to avoid excluding the reader from the group who is 'in the know'. By avoiding explicitly naming the agent, the writer can leave open whether the reader is part of the group who possesses the knowledge or not.

5.2.3 Active voice transitive patterns with complement clauses

Active voice transitive patterns involving complement clauses were significantly and meaningfully more frequent in the SNA corpus. As will be shown below, these patterns were often used to report people's mental processes and communicative acts, as well as interpret study findings.

Looking first at those transitive patterns with finite complement clauses, these patterns came in three varieties, namely patterns with *that* clauses following the main verb, patterns with *wh*-clauses following the main verb, and patterns with complement clauses preceding the main verb. Figure 5.5 shows that *that* clauses were heavily preferred in both registers, though the SNA corpus often positions the clause prior to the main verb of the sentence, also known as a 'reporting clause'.

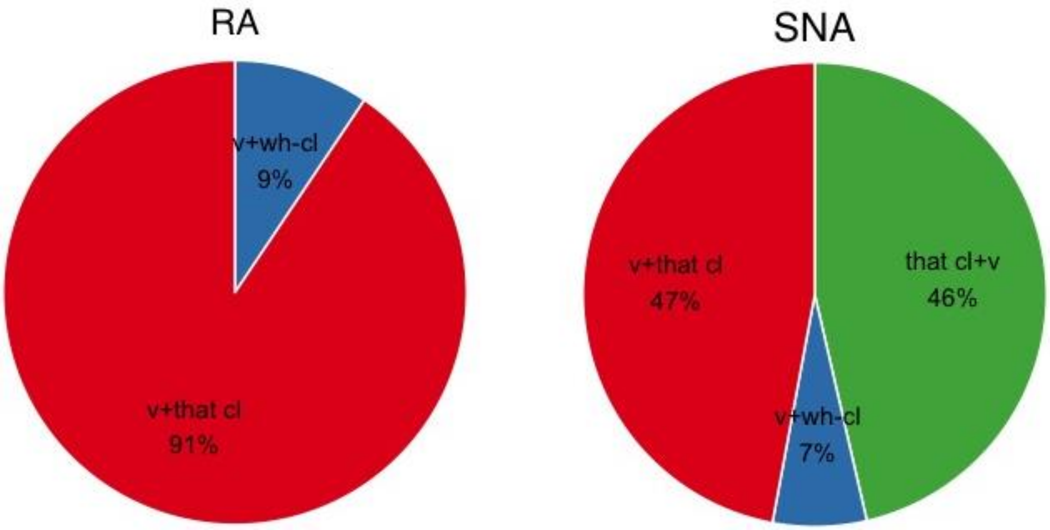


Figure 5.5 Two pie charts illustrating the proportional use of three varieties of transitive verbs complemented by clauses

These verb patterns functioned largely to report people's mental processes and communicative acts, as demonstrated by their most frequent subjects and verbs. Many of the most common subjects in the SNA corpus referred to humans, including *researcher, scientist, team*, *astronomer*, *we,* and *they*. Although the RA corpus also used animate subjects here (e.g., *patient*, *individual*, *we*), such sentences were less frequent overall (see Figure 5.1). Given that mental processes and communicative acts are generally attributed to humans, the presence of these patterns a focus on telling human stories in SNAs. Indeed, several informants confirmed this focus in our interviews:

> *When you're writing for popular audiences, it's easier to write in a more narrative sense. I think it is the difference between writing like a story, even if it's a news story, you're writing a story that's supposed to carry people along, and tell them something, versus a research article, which at its most basic you're recording exactly what you did. There is not much thought given to the actual enjoyability of the reading.* [Informant #6]

This excerpt illustrates that telling stories about research, rather than simply reporting it, may be an element of good SNA writing because it is more engaging for the reader—it carries the reader along and increases enjoyability.

The most frequent verbs in patterns with finite complement clauses also suggest a focus on people, including their communication (e.g., *suggest*, *say, show, note, indicate*) and cognition (e.g., *find, think, know, assume, believe*). Verbs of communication were particularly salient in the SNA corpus, owing to the larger presence of sentences where the reported clause comes before the subject and verb (5.12)-(5.13) (verbs **bolded**, finite complement clauses underlined):

(5.12) "It's not a bad thing to drink milk," Tucker **said**. [Health SNA #148]

(5.13) Scott’s chromosomal changes were likely exacerbated by high-energy particles and cosmic rays in space, Snyder **says**. [Health SNA #164]

These two excerpts demonstrate the traditional reporting clause syntax found in news reportage, where the reported speech comes first followed by the subject and main verb. Excerpt (5.13) also shows that this verb pattern is not only used for direct quotations, as in (5.12), but also paraphrasing the reported person’s words, as in (5.13).

When the finite clause comes in the more common post-verbal position, the main verb often describes mental processes rather than communicative acts. The flexibility offered by the finite complement clause allows the writer to describe the mental state in detail, as excerpts (5.14) and (5.15) demonstrate (verbs **bolded**, finite complement clauses underlined).

(5.14) Astronomers simply **do**n’t **know** exactly how these elements were able to make it into a gaseous form. [Space SNA #156]

(5.15) Moreover, we **hypothesized** that the finish time could be predicted by runners’ ability to anticipate their final time. [Psychology RA 215]

In short, transitive patterns with finite complement clauses were often used to attribute speech or thought to human subjects, and this purpose was particularly frequent in SNAs, demonstrated both by the greater density of subjects and verbs used to fulfill this function as well as the higher rates of these patterns overall.

However, not all active voice transitive patterns with finite complement clauses served to report people’s mental and communicative acts. For example, sentences with inanimate subjects and mental verbs were often used to offer an interpretation of the content of the complement clause. Consider the following two excerpts (verbs **bolded**, subjects underlined):

(5.16) This **means** researchers' understanding of cosmic collisions are usually obtained from laboratory simulations, and not first-hand observations. [Space SNA #217]

(5.17) This **could imply** that a more extreme spot contrast is justified at the center of the spot, where one might expect the umbra to be. [Space RA #91]

While earlier patterns attributed speech or thought to people, backgrounding the role of the writer, sentences like (5.16) and (5.17) allow the writer to comment on what the study and its components mean, imply, suggest, and more. For SNAs, this function arguably makes the article more valuable to readers, by going beyond a simple presentation of facts, quotations, and events.

Transitive patterns with non-finite complement clauses were considerably less frequent than their finite counterparts. There were two types of non-finite complement clauses, namely infinitive *to* clauses and *-ing* clauses, and both were relied on to a similar extent by RA and SNA writers (Figure 5.6).

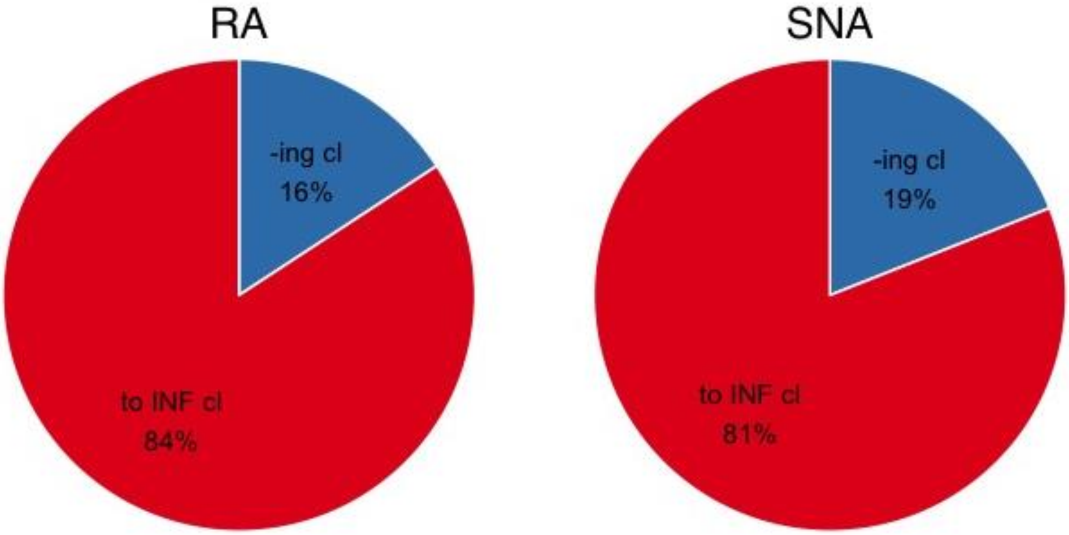


Figure 5.6 Two pie charts illustrating the proportional use of two varieties of transitive patterns with non-finite complement clauses

Figure 5.6 shows that both registers largely relied on *to* infinitive clauses to employ TV+NfCl patterns, likely due to the variety of verbs which can control these clauses (Biber et al., 1999). For this reason, Table 5.4 presents the most common verbs in TV+NfCl patterns by semantic domain rather than rank and frequency, as has been the standard procedure thus far.

Table 5.4 The ten most frequent verbs in TV+NfCl verb patterns across the registers organized by semantic category

Semantic Category	RA	SNA
	verb	verb
Aspect	continue, begin	start, begin, continue, go on
Communication	report	
Causation	help	help
Desire	want	want, hope
Effort	seek, fail	try
Intension	aim, choose	plan
Probability	tend	tend

Table 5.4 shows that the most common verbs in this pattern were semantically diverse. Among the categories more common in the SNA corpus were aspectual verbs and verbs of desire. Aspectual verbs tended to report research procedures, particularly when the subject referred to a specific researcher or study (5.18). When the subject referred to a collection of individuals, the sentence reported a changing understanding in a broader community (5.19) (verbs **bolded**, non-finite complement clauses underlined).

(5.18) To study that gradual transition, the researchers **began** <u>feeding mice high-fat mouse chow</u>, while periodically using a sophisticated microscope to look at the glutamatergic cells' ability to fire off signals. [Health SNA #141]

(5.19) Scientists **continue** <u>to discover more and more about the remote wilderness of Antarctica</u>. [Environment SNA #308]

Patterns with the verbs of desire *hope* and *want* often reported what scientists desired to accomplish in future research, stepping outside the confines of a study's immediate findings, as in (5.20) (verbs **bolded**, non-finite complement clauses <u>underlined</u>):

(5.20) Astronomers, including Bodewits and Bromley, **are** now **hoping** <u>to comb through other archives to see if they can find additional evidence for metals, or other interesting results</u>. [Space SNA #156]

Sentences like (5.20) suggest that SNAs may include a function like that of "future directions" found in many RAs. Indeed, one informant noted that future directions are something that they often include at the end of their articles:

> *Because we really want to emphasize for readers that no single paper is like the be-all-end-all in any field of science, towards the end of the study we'll talk about, like, what were the maybe the methodological limitations, what future research still needs to be done, what lingering questions are left.* [Informant #5]

In the RA corpus, verbs of effort (*aim, seek*) and intension (*fail, choose*) were relatively common. *Aim* and *seek* stated or re-stated the goals of a study, while *fail* reported whether that goal was achieved or indicated a gap in the literature. *Choose* reported the authors' deliberate choice to follow a particular procedure.

Help, by contrast, was the most common verb of TV+NfCl patterns in both registers. It served an interpretive function in discourse. That is, writers used these patterns to interpret the importance or impact of a study's results by connecting subjects like *findings* or *study* with what they could *help* people accomplish in the future. Consider excerpts (5.21) and (5.22) below (verbs **bolded**, non-finite complement clauses underlined):

(5.21) The finding, published in the April 12 Science, **may help** explain how the hallucinogenic anesthetic can ease some people's severe depression. [Health SNA #145]

(5.22) Importantly, our study **can help** to optimize future mapping efforts, and fill data gaps where information is lacking. [Environment RA #355]

It is notable that both (5.21) and (5.22) include a modalized verb phrase to interpret the utility of the study's findings, likely used to hedge the proposed interpretation. Modal verbs will be discussed in more detail in Chapter 6.

5.3 Copular verb patterns

Section 5.2 highlighted a number of frequent verb patterns that were used with significantly different frequencies between the two registers. This section turns to another group of patterns that were relatively common but showed less severe frequency differences between the registers, namely verb patterns with copular verbs. Three verb patterns with copular verbs were shown in Figure 5.2, namely copular patterns with phrasal complements (CV+PhC), copular patterns with non-finite clausal complements (CV+NfCl), and copular patterns with finite clausal complements (CV+FCl). Below, I examine these patterns in the contexts of use.

The three copular verb patterns mentioned above, namely CV+PhC, CV+NfCl, and CV+FCl, can be further specified into seven more specific verb patterns by the nature of the phrasal complement and the nature of the complement clause. Figure 5.7 illustrates the proportional use of these fine-grained verb patterns.

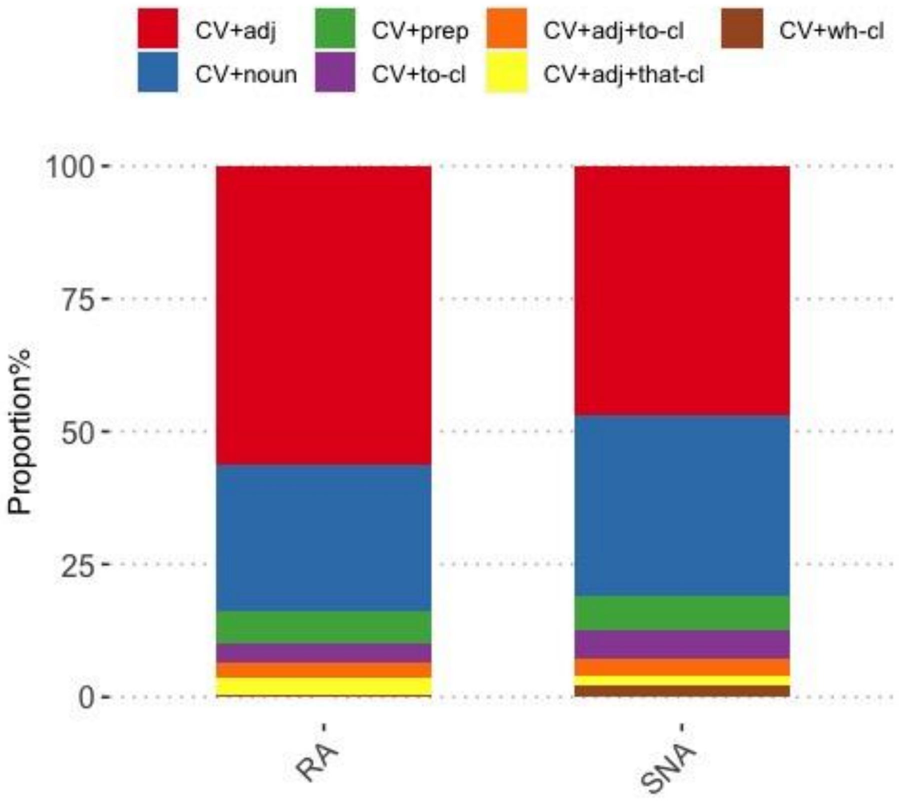


Figure 5.7 A bar graph illustrating the proportional use of seven varieties of verb patterns with copular verbs across registers

Figure 5.7 shows that most copular patterns had adjective phrases or noun phrases as post-verbal complements, reflecting the much greater frequency of the CV+PhC pattern relative to other copular patterns (see Figure 5.1). RAs relied most on adjective phrase complements while SNAs relied more on noun phrase complements. The remaining five patterns, which largely include clausal complements, were infrequent by comparison. For example, a prototypical RA or SNA uses less than one CV pattern with a clausal complement for every five CV patterns.

Below, I first focus on the similarities and differences in the use of CV+noun and CV+adj verb patterns across the registers, before briefly examining CV patterns with clausal complements.

To gain deeper insight into the use of copular patterns with noun phrase complements, I first examined them for whether the noun was determined by the definite article *the*, indefinite article *a/an*, or something else (i.e., the zero article or a non-article determiner). This examination revealed that noun complements of CV patterns in the SNA corpus were significantly associated with the indefinite article (adjusted residual = 3.65), while the RA corpus was statistically associated with the 'other' category (adjusted residual = 4.39) ($\chi^2(2)=21.12$, p < .001). However, the overall effect of this association is small (Cramer's *V* = 0.11). In short, CV + *a/an* noun patterns were more descriptive of SNAs, while CV+nouns determined by ø or a non-article determiner were more descriptive of RAs.

The positive association of *a/an* with CV+noun patterns in the SNA corpus suggests a greater presence of singular countable nouns there relative to the RA corpus. By contrast, many of the most frequent nouns in CV+noun in RAs referred to measurements, such as *year*, *mm*, and *percent*, where the noun was determined by a numeral rather than an article. Excerpts (5.23) and (5.24) illustrate these features (copular verbs **bolded**, noun phrase complements <u>underlined</u>):

(5.23) Full genome sequencing **is** <u>a torturously long process</u> - but a lot of the time, there are only specific parts of the genome you're interested in. [Health SNA #231]

(5.24) The associated average estimate of surface accumulation for the last thousand years **is** <u>~47 mm</u>. [Environment RA #307]

Excerpt (5.23) represents a typical instance of CV+*a/an* noun in the SNAs, where the noun is singular, countable, and used to define and characterize subjects using stance-laden language like *tortuously long*. In excerpt (5.24) from an RA, the noun complement *mm* instead is determined by the numeral *~47* and thus is not determined by *a/an*.

Similarly, differences were found in the most frequent adjectives used in CV+adj patterns between the registers. In the RA corpus, *consistent* was the most frequent adjective in this position, occurring 500 times (0.25 ptw), while the same adjective was used just ten times (0.04 ptw) by SNA writers. In the SNA corpus, *able* was the most frequent adjective in this position, occurring 78 times (0.28 ptw), while the same adjective was used just 148 times (0.07 ptw) by RA writers. Excerpts illustrating these characteristics are shown below (copular verbs **bolded**, adjective complements <u>underlined</u>):

(5.25) This finding **is** <u>consistent with previous literature that suggests that anxiously attached individuals harbour chronic relationship doubts with regard to their partner's love and acceptance</u>. [Psychology RA #189]

(5.26) By combining the two technologies, the team **was** <u>able to get a better picture of the Dark Zone than ever before</u>. [Space SNA #326]

As excerpt (5.25) shows, CV+adj patterns with *consistent* often relate one study referred to by the subject with other research in the field referred to by the noun phrase complementing *consistent with*. In these sentences, researchers evaluate the findings of their own study by comparing them with previous work, contextualizing their new work within a greater scholarly debate. By contrast, CV+*able* patterns like (5.26) tended to have human subjects and described their ability to accomplish some task. Thus, these sentences allow writers to narrate science by portraying research as a process of humans planning, attempting, and succeeding in goal-oriented activity.

Finally, turning to copular patterns with complement clauses, Figure 5.8 displays the mean normalized rates of four different fine-grained verb patterns across registers.

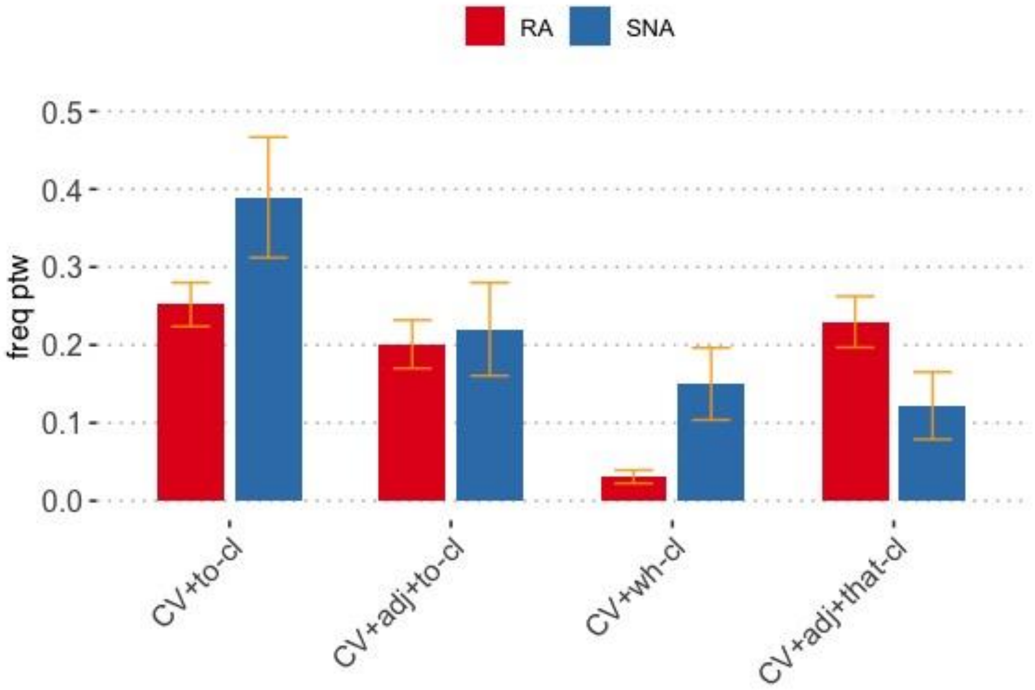


Figure 5.8 A bar graph illustrating the mean normalized rates of copular patterns involving complement clauses

The 95% CIs shown in Figure 5.9 suggest that CV+*to*-cl and CV+*wh*-cl were significantly more frequent in the SNA corpus, while CV+adj+*that*-cl patterns were significantly more frequent in the RA corpus. CV+*adj*+*to*-cl patterns were used with similar frequency between the registers. In general, all differences were marginal, with the most meaningful differences ranging from a near-medium (CV+wh-cl, *d* = 0.41) to small (CV+*to*-cl, *d* = 0.23) effect size.

The primary reason for the greater use of CV+*to*-cl patterns in the SNA corpus is the greater occurrence of these patterns with the subject *step*. An example is shown in excerpt (5.27) (copular verb **bold**, non-finite complement clause underlined, subject head *italicized*).

(5.27) The next *steps* for the team **are** to conduct clinical trials, and possibly investigate whether other molecules should be targeted as well. [Health SNA #200]

Excerpt (5.27) demonstrates that the subject *step* is often used to identify future research projects extending beyond the immediate one being summarized, similar to the function shown in excerpt (5.20). Again, a focus is placed on human cognition (i.e., planning) and its relation to further actions (i.e., conducting future investigations), serving to turn the scientific process into a story of science. Naturally, sentences like (5.27) are often found at the end of the article (cf. Nwogu, 1991, p. 118).

Like CV+*to*-cl patterns, CV+*wh*-cl was also significantly more frequent in SNAs. The most frequent variety of this pattern was a nominal relative clause with the subject *that*, while the most variety in the RA corpus was a dependent interrogative clause with the subject *question*. Excerpts (5.28) and (5.29) illustrate these characteristics (verbs **bolded**, *wh* complement clauses underlined, subject heads *italicized*).

(5.28) *That*'**s** why scientists had long assumed that geologic features like craters, even if they once existed, would smooth out over time if they were under the ice. [Space SNA #264]

(5.29) Another open *question* **is** whether givers may give a gift to soften a future transgression. [Psychology RA #285]

Sentences like (5.28) highlight answers to questions, i.e. they link the answer referred to by *that* with the question-like *wh* clause. The fact that these sentences often have the pronoun *that* as their subject also points toward the writers' attempt at building cohesive links between units of discourse (Padula et al., 2020). *That* points back to something prior in the text, which the writer then uses as an initializing theme of the new sentence and is able to comment upon that theme with the complement clause. By contrast, sentences like (5.29), which were more common in RAs, present an indirect question without an answer—there remains an outstanding *key question* that may or may not be answered by the study at hand. As a result, sentences like these were often found near the beginning of articles, serving to indicate a gap in the literature, or near the end, serving to indicate potential areas of future research.

By contrast, verb patterns of the form CV+adj+*that*-cl were more frequent in RAs. Many instances of this pattern have the semantically empty *it* as its subject, sometimes called a 'dummy *it*' or 'anticipatory *it*'. This pattern has been shown to express writer stance toward the content of the *that* clause (Biber et al., 1999). Moreover, because the writer his or herself is not explicitly mentioned in the sentence, the pattern expresses stance in a particularly impersonal way preferred by academic writers (Hewing & Hewing, 2002). While the CV+adj+*to*-clause used a variety of semantic groups of adjectives (e.g., *important, hard, urgent*), CV+adj+*that*-cl tended to use adjectives of likelihood, especially *possible* (5.30) (verb **bolded**, adjective *italicized*, *that* complement clause underlined).

(5.30) It **is** *possible* that the differences found among knowledge conditions may be an artefact of poor comprehension rather than actual differences in knowledge attributions. [Psychology RA #153]

There are at least two potential reasons for the greater use of this verb pattern in RAs. First, as mentioned before, they represent a particularly impersonal and academic style of expressing stance. SNA writers may not feel the same pressure to disguise personal stance in such a way.

Second, given the fact that this dissertation's SNAs have been shown to rely on verb patterns serving to report others' words and ideas (see section 5.2), there may simply be less of a need, or less room, for SNA writers to express their stance in general, leading to fewer instances of CV+adj+*that*-cl overall.

5.4 Cluster analysis

Thus far, I have considered a handful of verb patterns individually to examine how they are used by RA and SNA writers in texts. In this section, I broaden the scope of analysis by considering multiple patterns at once and look at longer stretches of discourse, accomplished by adopting cluster analysis to group texts together based on their use of verb patterns. As noted in section 3.1.5.3, the top ten most frequent verb patterns across the registers were included in the cluster analysis to avoid including too many variables for the number of observations. Specifically, the verb patterns TV+PhO, CV+PhC, TV+FCl, FCl+S|TV, Pass TV, TV+NfCl, IV+PhC, ExThere+BE, CV+NfCl, and CV+FCl were included. Additionally, two SNAs were excluded from the analysis because they were found to exert undue influence on the clusters. Thus, 798 total texts (398 SNAs and 400 RAs) were analyzed.

5.4.1 A three-cluster solution

A dendrogram was produced and visually inspected to determine the optimal number of clusters (Figure 5.9).

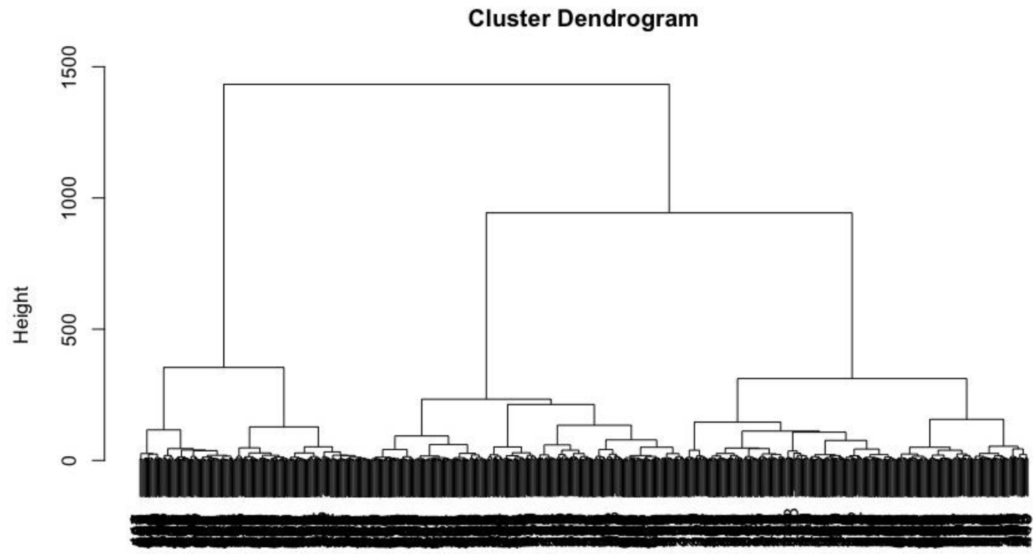


Figure 5.9 A dendrogram illustrating the clustering of 798 RA and SNA texts into groups at different levels

This dendrogram lists the texts along the bottom of the figure. Similar texts are grouped more closely to one another, and lines indicating the relative similarity between groupings are drawn to link those groups or clusters together. Significant gaps between banks of horizontal lines suggest that a particular grouping of clusters minimizes the dissimilarity between clusters (as quantified on the y-axis) while also identifying those texts that are similar to one another. For example, Figure 5.9 suggests that a three-cluster solution is the most optimal one for grouping the 798 SNAs and RAs, since a horizontal line drawn across unit ~750 of the y-axis neatly classifies all texts into three groups.

With these three clusters identified, I then examined the composition of these clusters in terms of the register and topical domain of the texts within them. Of the three clusters, one is largely comprised of SNAs, another of RAs, and another of either register (Figure 5.10).

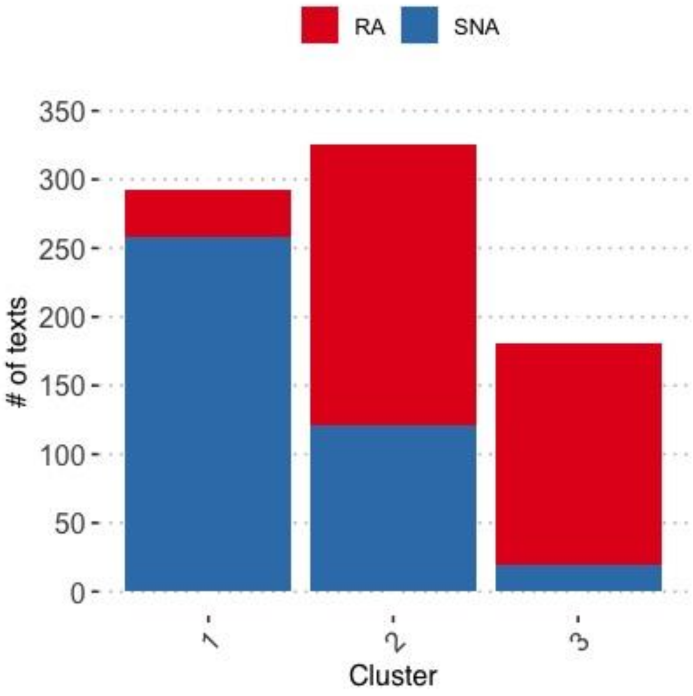


Figure 5.10 A bar graph showing the number of texts from each register by cluster

Figure 5.10 shows that Cluster 2 had the largest number of texts, likely since this cluster included many texts from both registers. Cluster 1 had just under 300 texts, mostly SNAs, and Cluster 3 showed the fewest number of texts, of which most were RAs.

With regard to the topical domains of texts, all three clusters used similar proportions of environment and space texts, while the proportions of health and psychology texts changed more significantly from Cluster 1 through Cluster 3 (Figure 5.11).

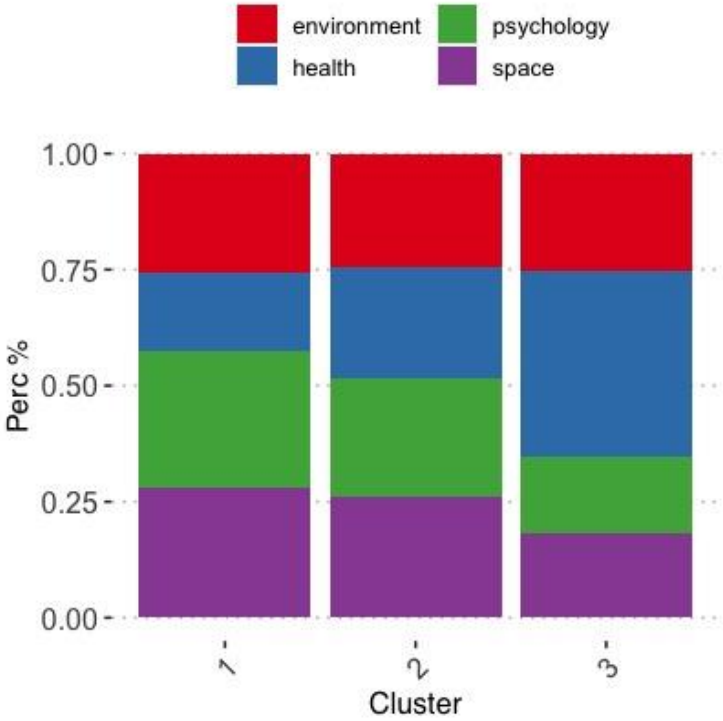


Figure 5.11 A proportional bar graph illustrating the percentages of topical category by cluster

Figure 5.11 shows that psychology and space texts tended to decrease from Cluster 1 to Cluster 3, while health texts increased. This trend is most notable for health texts, which contributed substantially to Cluster 3 but minimally to Cluster 1. Additionally, Figure 5.11 also shows that environment texts remained roughly the same across the clusters.

With these basic descriptions of the clusters, I next (5.4.2) turn to interpreting the meaning of the three clusters by describing the verb patterns that comprise them, as well as provide longer excerpts to illustrate what their discourses look like.

5.4.2 Interpreting the three clusters

To understand how the ten verb patterns relate to the three clusters, mean normalized frequencies for the patterns were calculated for each cluster. Each cluster was then defined by the patterns which were most and least frequent in that cluster (Figure 5.12).

Cluster 1	+	CV+PhC, CV+NfCl*, Ex*There*+*BE*, FCl+S|TV*, TV+FCl*, TV+NfCl*
	−	IV+0, PassTV*, TV+PhO*
Cluster 2	+	TV+PhO*, IV+0
	−	Ex*There*+*BE*, PassTV+NfCl*
Cluster 3	+	PassTV*, PassTV +NFCl*
	−	CV+PhC, TV+FCl*, FCl+S|TV*, TV+NfCl*, CV+NfCl*

*Significantly different in this cluster relative to one or both of the other clusters, based on 95% CIs

Figure 5.12 The most (+) and least (−) frequent verb patterns in each cluster

Figure 5.12 shows that Cluster 1, which had the highest proportion of SNAs, is defined by the presence of complement clauses and copular patterns, as well as a low frequency of passive transitive verbs, transitive verbs with phrasal objects, and intransitive verbs. By contrast, Cluster 2 is defined by the presence of transitive verbs with phrasal objects and intransitive verbs, as well as the lack of existential *there* patterns and passive transitive verbs with non-finite clause complements. Finally, Cluster 3, which had the highest proportion of RAs, showed high rates of patterns involving passive voice transitive verbs, as well as low rates of transitive and copular patterns with clausal complements. Below, I describe these clusters, beginning with Cluster 3, and provide excerpts for each.

Beginning with Cluster 3, this cluster is characterized by the presence of RAs, especially health RAs. For these texts, the passive voice is used to front the object of the transitive verb and (usually) omit the agent. Consider the following excerpt from a health RA (passive voice verb phrases **bolded**):

> **Preventing post-surgical cardiac adhesions with a catechol-functionalized oxime hydrogel**
>
> A variety of materials **have been studied** for the reduction and prevention of surgical adhesions, such as Seprafilm (sodium hyaluronate/carboxymethylcellulose sheet) and CoSeal (poly-ethylene glycol (PEG) hydrogel)[10-15]. However, limited success **has been demonstrated** for preventing or reducing the severity of cardiac adhesions due to short retention time as a result of dynamic motion of the heart, short degradation times, and excessive swelling of the polymer leading to cardiac tamponade[10-12,14-18]. One product **was approved** in the US for preventing cardiac adhesions, REPEL-CV (polylactic acid/PEG sheet), although this failed to reduce adhesion dissection time[19] and is no longer sold.
>
> [Health RA #140]

The main verb in all three sentences of this excerpt is passive. As a result, the sentences emphasize objects, such as *materials* and *products*, and outcomes, such as *limited success*. In the first and second sentences, the implied subjects are previous publications (i.e., [10-15], [10-12,14-18]), while the third sentence implies the Federal Drug Agency as the agent. Academic writing is often concerned with making generalizations (Biber et al., 1999, p. 938-939), and by stating propositions in the passive voice, academic writers convey propositions as generalizable rather than the idiosyncratic result of particular researchers. The RA writer of the excerpt above, for example, uses the passive voice to portray a problem in surgical materials as extensive and significant.

Of course, passive-heavy writing is perhaps most infamous for its use in the methods sections of RAs, where the agent is usually the scientist and thus already known to the reader (passive voice verb phrases **bolded**):

> **High-resolution cryo-EM structures of outbreak strain human norovirus shells reveal size variations**
>
> For asymmetric focused reconstruction, a 3D binary mask of an ASU **was generated** using Chimera and IMAGIC (33, 34). An inverse version of this mask **was** then **applied** to the symmetrical 3D reconstruction using apply_mask within cisTEM. This masked reconstruction **was used** to subtract out all but one ASU from the raw images in combination with symmetry expansion. For symmetry expansion, each image **was subtracted** 60 times using each of the 60 icosahedral ASUs to create a dataset of isolated asymmetric units, which were automatically centered.
>
> [Health RA #179]

The usefulness of the passive voice in methodological description is fairly straightforward. Readers intuitively understand that the agent(s) of the methodological steps is the same researcher(s) who authored the previous sections of the article that they have already read, and thus repeated mention of that agent is not essential. Nonetheless, omitting the agent and thematizing the direct object has other discoursal consequences. For example, I have already shown that some SNAs writers see the purpose of science news writing as telling stories involving human actors. Thus, a heavy use of passive voice would be at odds with this purpose.

However, not all RAs were grouped into Cluster 3. As Figure 5.10 and Figure 5.11 illustrated, many RAs and SNAs were grouped together into Cluster 2, defined by, in particular, the frequent use of transitive verbs with phrasal complements. Thus, while Cluster 3 was defined by the promotion of phrasal objects to grammatical subjects, in Cluster 2 those objects remained in their post verbal position, resulting in a greater presence of explicit agents in the subject position. To get a sense of the discourse of Cluster 2, consider the following excerpt from a psychology RA (TV+PhO VPs **bolded**):

> **Candidate aggression and gendered voter evaluations**
>
> In the American context, South Carolina Republican Representative Joe Wilson **gained** notoriety for yelling "You lie!" at President Obama, and Republican Montana congressional candidate Greg Gianforte **body slammed** a reporter in 2017. Women **are getting into** the fighting spirit too. Democratic House Speaker Nancy Pelosi **received** criticism for incivility when she tore the text of President Trump's 2020 State of the Union address. Incivility is hardly unique to American politics. During the 2017 French Presidential election, Marine Le Pen, **called** her opponent "cold-eyed," "arrogant," and "spoilt."
>
> [Psychology RA #445]

In this excerpt, which comes from an introduction section, the writer narrates past events involving political incivility, resulting in several past tense active voice transitive verbs with phrasal objects. However, such narratives were relatively infrequent in RAs. Instead, the most apparent reason for the inclusion of many RAs in Cluster 2 is the willingness to use active, rather than passive, verbs, especially in methodology sections. Consider the following excerpt (TV+PhO verbs **bolded**):

> **A relation between the radial velocity dispersion of young clusters and their age**
>
> First, for the profile fitting, we **adopted** Gaussian profiles. Second, we **clipped** the core of diagnostic lines that were still contaminated by residuals of the nebular emission. Third, we simultaneously **fit** all spectral lines available, thereby assuming that the Doppler shift is the same for all lines. Figure C.1 **shows** the radial-velocity distribution for the two regions. We **calculated** the errors of the histogram bins by randomly drawing RV values from a Gaussian centered at each measured RV and with a sigma corresponding to the measurement error.
>
> [Space RA #52]

The use of the subject *we* in the above excerpt reflects the willingness of some RA writers to use the active voice to narrate the methodological steps. When an RA's writer used active voice to describe methods, the overall rate of passive voice verbs decreased, resulting in a greater chance of clustering with Cluster 2 rather than Cluster 3. Thus, despite the stereotypically impersonal style of academic writing, some writers, or more accurately some journals (see, e.g., Millar et al., 2012), are willing to inflect the scientific method with the subjectivity of individual researchers by explicitly mentioning themselves via first person pronouns in their prose. Some SNA writers also wrote with these linguistic features (TV+PhO verbs **bolded**):

Mild zaps to the brain can boost a pain-relieving placebo effect

> Participants **reported** lower pain intensity from the heat on the “lidocaine” patch of skin, an expected placebo effect. People also **reported** higher pain intensity on the “capsaicin” skin, an expected nocebo effect. Before testing the placebo and nocebo effects, researchers **had delivered** electric currents to some participants’ brains with a method called transcranial direct current stimulation, or tDCS. During these tDCS sessions, two electrodes attached to the scalp **delivered** weak electric current to the brain to change the behavior of brain cells. Some participants **received** tDCS targeted at a brain area thought to be important in placebo and nocebo effects, the right dorsolateral prefrontal cortex. Researchers **used** two types of current: positive anodal tDCS, which typically makes nerve cells more likely to fire off signals, and negative cathodal tDCS, which usually makes cells quieter.
>
> [Health SNA #19]

The six bolded verbs are past tense, active voice, and transitive. Several verbs have human subjects (*participants, researchers*) and report methodological actions (*deliver, use*). SNAs in Cluster 2 share these features. That is, they piece apart the study’s methodological steps, recounting them sentence-by-sentence, and report their outcomes. The result is a discourse style in which the writer presents an accurate re-telling of what the researchers did to reach their findings and what those findings were. Thus, in this part of Cluster 2 SNAs, the focus is less about implications and applications, which often occur with the present tense or modal verbs, and more about methods and findings.

The final cluster, Cluster 1, is similar to the discourse style of Cluster 2 in that they both utilize active voice transitive verbs. However, Cluster 1 is defined by the higher frequency of transitive verbs complemented by clausal, rather than phrasal, complements, often reporting the words and cognitive states of humans (see section 5.2.3). Thus, the FCl+S|TV, TV+FCl, and TV+NfCl patterns were most frequent in Cluster 1. In addition, a pair of copular verb patterns were also more frequent here (i.e., CV+PhC, CV+NfCl). To unpack this cluster, consider the following excerpt from an SNA (transitive verbs complemented by clauses **bolded**, copular verbs underlined):

> **Babies Are Predicted to Be Born Earlier in The Extreme Heat of Climate Change**
>
> Compared to full-term pregnancies, we **know** that near-term babies have a higher risk of medical problems soon after birth, and lower cognitive outcomes later in childhood. So, with that in mind, we have some bad news. Researchers have **found** that extreme heat makes babies rush to the exit sooner, leading to an average of 25,000 US infants a year born a little early due to hot weather. And like nearly everything else in this world, it's only going to get worse with climate change. "Given recent increases in the frequency of extremely hot weather, there is a clear need to better forecast the potential magnitude of climate change's impact on infant health at the national level," the team **explains** in their paper.
>
> [Environment SNA #314]

Of the five sentences in this excerpt, three sentences include a TV+FCl pattern, serving to attribute knowledge (*we know that…*) and words (*"…" the team explains*) to people, and one includes a CV+PhC pattern, which links a stance adjective to its subject (*it…worse*). As this excerpt suggests, transitive verb patterns with finite complement clauses often serve to report others' words or thoughts, while the copular patterns found in Cluster 1 define and characterize (usually) noun phrase subjects. In doing so, these copular verb patterns allow the SNA writer a chance to express their stance, define important concepts for readers, and build cohesion between sentences otherwise focusing on the thoughts and words of people other than the writer. The following excerpt from a space SNA makes this role of copular verb patterns more clear (copular verbs **bolded**):

> **Aliens could exist on worlds weirdly similar to a classic Star Wars planet**
>
> Binary star systems may **seem** more sci-fi than reality, but in truth, they **are** very common in our galaxy. Roughly half of Sun-like stars in our neighborhood **are** in binary systems. Knowing whether these systems are hospitable to life **is** a major component in our search for life in the universe beyond our own planet.
>
> “It**’s** very interesting to know that around the stellar binaries. Even with the influence of the gas giants, many terrestrial planets would still be quite stable,” Gongjie Li tells *Inverse*. Li **is** an astrophysicist at the Georgia Institute of Technology who was not involved in the study.
>
> [Space SNA #48]

This excerpt is brimming with copular verb patterns. In fact, each sentence includes one, though the penultimate sentence is counted as a FCl+S|TV verb pattern since *tells* controls the preceding sentence (i.e., *Even with*…). These sentences communicate stance, (*may seem*, *very common, very interesting*), locate objects in space (*roughly half…are in binary systems*), and characterizes ideas (*knowing whether…is a major component*) and people (*Li is an astrophysicist*). By performing these functions, copular verb patterns in SNAs cohesively tie in direct and indirect reports of scientists' words and ideas, striking a balance between having a purely reporting function and calling on no outside sources at all.

5.5 In focus: Optional adverbials across register and topical category

So far, I have focused on the required elements in verb patterns, such as subjects, verbs, and objects. In this section, I examine a grammatically optional element, namely optional adverbials (hereafter, just 'adverbials'). Adverbials largely come in the form of adverb phrases (e.g., *still*), noun phrases (e.g., *today*), prepositional phrases (e.g., *in the journal Nature*), and clauses (e.g., finite concession clauses, non-finite purpose clauses). A linear mixed effects model was fit onto the data to identify whether the use of adverbials varied as a function of register, topical domain, or an interaction of both. The model of best fit reported significant main effects for register and domain, but no interactional effect, with a moderate effect size (R2c = .39) (see Appendix F.1 for the full output). Specifically, adverbials were significantly more frequent in RAs, significantly more frequent in health texts, and significantly less frequent in psychology texts. These findings are explored below.

RA writers used more adverbials on average than SNA writers (41.7 ptw to 35.4 ptw, respectively), largely due to prepositional phrases functioning as adverbials (Figure 5.13).

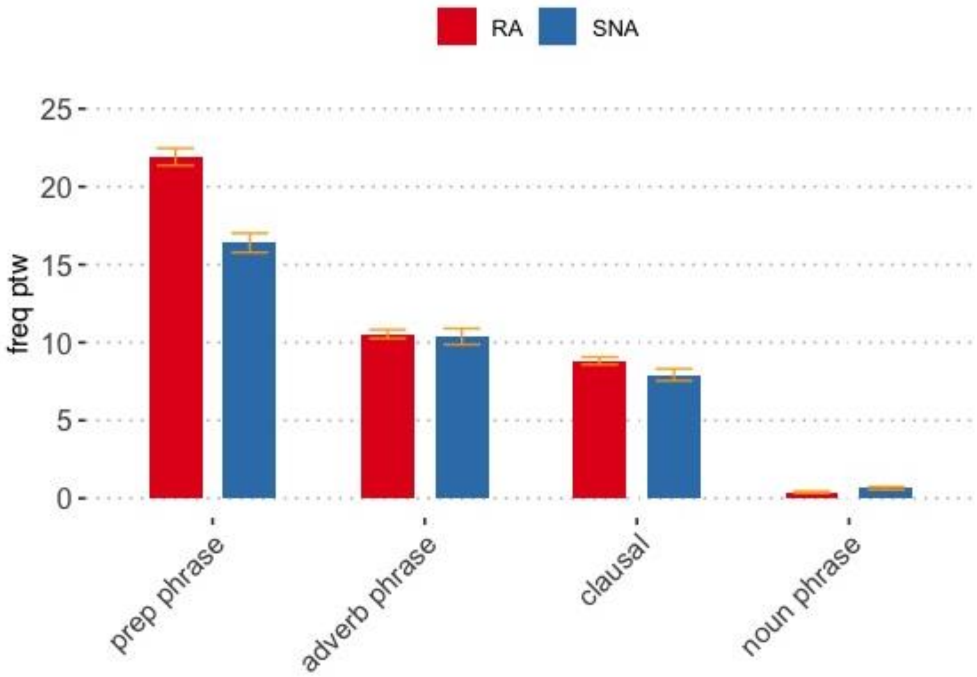


Figure 5.13 A bar graph illustrating the mean normalized rates of four types of adverbials across the registers

The higher rate of adverbials in RAs is arguably due to two reasons. First, adverbials were particularly frequent in PassTV and TV+PhO patterns, both of which were more frequent in the RA corpus (see Figure 5.1). Second, adverbials were relatively *in*frequent in TV+FCl and FCl+S|TV patterns, since any adverbials tended to be placed within the complement clause rather than main clause. Thus, patterns with higher rates of adverbials were more frequent in the RA corpus, while patterns with lower rates of adverbials were more frequent in the SNA corpus.

Topical domain also played a factor. Health texts showed the highest rate of adverbials in both registers, while psychology texts used the fewest (Figure 5.14).

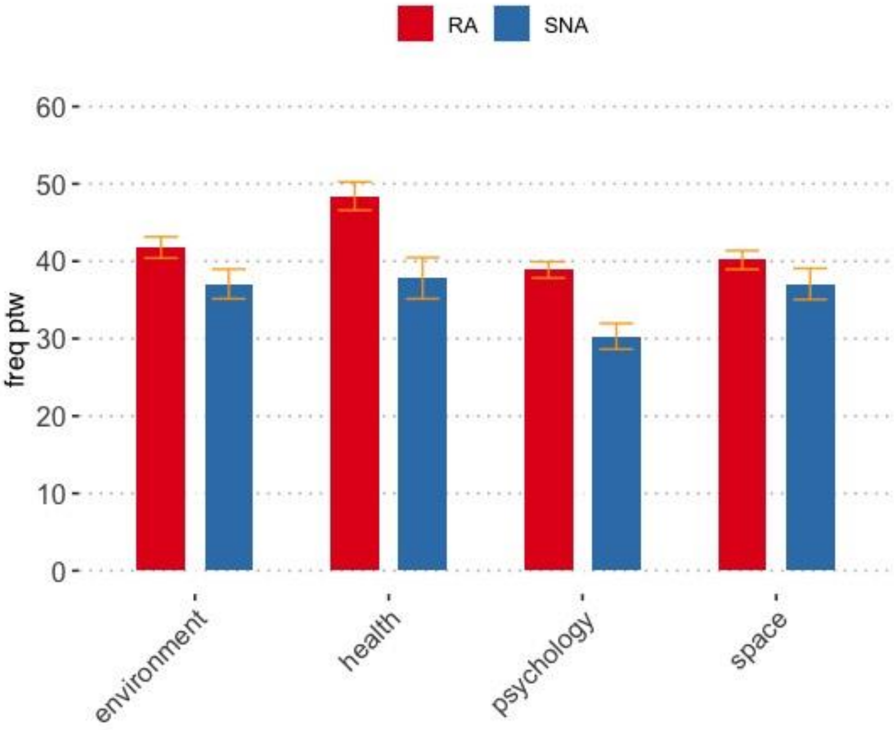


Figure 5.14 A bar graph illustrating the mean normalized rates of adverbials by corpus and topical category

Figure 5.14 shows that, in the RA corpus, adverbials were significantly more frequent in health texts than the other domains, while in the SNA corpus, frequency differences were less severe. Nonetheless, health SNAs still boasted the highest rate of adverbials among the four topical domains, suggesting a usefulness for this feature in RAs and SNAs communicating health research. Similarly, psychology SNAs used fewer adverbials than other SNAs, while psychology RAs also showed the fewest adverbials across the domains.

One reason for higher rates of adverbials in health texts is their higher rate of PassTV patterns, where they were used to describe methodological procedures with precise detail. However, a casual review of methods sections of non-health RAs also uncovered a dense use of adverbials, so the higher frequency of adverbials in health RAs must be the result of either longer methods sections (and thus more adverbials) or more adverbials used in other RA sections. For example, consider the following excerpt from the introduction of a health RA (adverbials underlined):

> **An organosynthetic dynamic heart model with enhanced biomimicry guided by cardiac diffusion tensor imaging**
>
> There is an unmet clinical need for a high-fidelity benchtop cardiac model for device testing, interventional training, and procedure demonstration. Currently, in most in vitro cardiac simulators, the heart is represented by entirely synthetic or entirely organic components. Although a realistic representation of cardiac anatomy is achievable by using ex vivo beating heart models, the setup process can be long and tedious, and the ex vivo heart tissue has limited longevity due to muscle stiffening and decay. Owing to these drawbacks, in vivo animal models are commonly used in industry to test the mechanical performance of intracardiac devices, involving substantial experimental cost and time.
>
> Health RA #62

In about 100 words, this excerpt shows nine adverbials of various forms, including adverb phrases (e.g., *currently*), prepositional phrases (e.g., *in industry*), and finite and non-finite clauses (e.g., *although a realistic…*). The distribution of these forms also roughly mirrors Figure 5.13, where prepositional phrases functioning as adverbials made up about half of all adverbials in the RA corpus. Syntactically, these phrases and clauses are optional, but they add important communicative detail. Adverbials in this excerpt describe causes, reasons, concessions, and results, making the writers' argument clearer and more explicit.

By contrast, psychology texts used the fewest adverbials, especially in the SNA corpus. While psychology texts tended to use all adverbial forms less frequently than the other topical categories, they used particularly fewer prepositional phrases and clauses functioning as adverbials. Consider the following excerpt from the start of a psychology SNA (adverbials underlined):

> **Study sheds light on how Black students' stereotypes about academic performance change as they age**
>
> A longitudinal study published in Developmental Psychology found that Black children possess different racial stereotypes in regard to academic and nonacademic performance. African American girls perceived Black children as less academically competent in STEM subjects compared to White children. However, Black children were perceived as more competent in music and sports. "These perceptions likely have significant consequences for the academic achievement of African American youth. Because of the importance of social identities for individual identity, youth are likely to pursue domains in which they perceive that other members of their own group excel," wrote the authors.
>
> Psychology SNA #248

This excerpt shows two adverbial, *however,* attributable to the writer, and *likely*, attributable to the quoted authors. Other adverbials are present; however, they are couched within a finite complement clause controlled by the verb of communication *wrote*, highlighting the relationship between the frequency of adverbials and the verb patterns involved. As a result, this discoursal style presents the writer as a conduit for reporting the facts of the study and the words of its researchers. The writer is neither presenting an argument nor interpreting findings but instead presenting the facts a given study. Interestingly, this style connects to the situational characteristics of its texts. That is, the lack of adverbials was particularly severe for one source, *The Academic Times*, which was also the only source to state its intended audience as being, in part, active researchers (see 4.1.7.1). Perhaps, then, these writers see their role as plainly providing the procedures and findings of recent research, which the reader, who themselves are also academics, can make their own interpretation about.

5.6 Conclusion

In this chapter, I compared the use of 30 specific verb patterns, grouped into 14 more inclusive patterns for easier visualization and comparison, across the RA and SNA corpora. Verb patterns describe the obligatory elements of a clause, such as subjects, verbs, and objects, and thus are like valency patterns (see Biber et al., 1999). Active voice transitive verbs with phrasal objects were the most frequent patterns in both registers, though they were significantly more frequent in RAs. RA writers also used many passive voice transitive verb patterns, particularly when describing research methods. The passive verb patterns often emphasized the activity of researchers, where main verbs were those like *use, perform*, and *calculate*, while the active patterns emphasized a wider range of semantic domains attributable to non-human subjects, with main verbs such as *have*, *show*, and *include*. By contrast, SNA writers used many patterns with complement clauses, especially finite *that* clauses such as reporting clauses, which place a direct or indirect report before the subject and main verb. These patterns emphasize the words and ideas of people other than the writer. Finally, copular verb patterns were used with similar frequency across the registers. In SNAs, these patterns were often used to express stance and make texts with many reporting clauses more coherent, making such texts more than simple summaries of their respective RAs.

Naturally, these findings can be linked to the situational characteristics of the registers. For example, SNAs are written by lay experts and not the researchers of the studies themselves. In addition, the locations in which SNAs are published resemble news outlets, with a staff of writers, editors, and new articles published at short, regular intervals. As a result, SNAs are brief, consumable in a short sitting with few interruptions by sections, tables, or figures. Writers adopt verb patterns that help them tell stories of science, which both make the writing more relatable and add credibility, since it is not solely the journalist who is summarizing and interpreting the research. By contrast, RAs are written by the same experts who conducted the research and are published in locations where, in large part, only other active researchers look. Moreover, the purpose of the research is not to entertain but to transform knowledge and provide enough detail and evidence to convince readers and allow for replication, if applicable. Thus, RAs are long, detailed texts with many sections and visuals to break up the reading experience. Writers employ verb patterns that emphasize their objects of study, usually inanimate objects, including the actions done upon them by researchers and their abstract relationships with other things, processes, and events. At the same time, neither register is completely homogenous, evidenced by the three-cluster solution described in 5.4.

6 VARIATION IN THE SHORT VERB PHRASE

In this chapter, I describe findings from analysis of the short verb phrase across the RA and SNA registers. Specifically, I seek to answer how tense, aspect, modality, and voice are used similarly and differently between the corpora, using corpus and informant data to inform the analysis. In section 6.1, I compare the rates of finite and non-finite verb phrases, followed by a comparison of tense and aspect in 6.2, grammatical voice in 6.3, and modal verbs in 6.4. Then, in section 6.5, I present a cluster analysis of these features, and in section 6.6 I examine how the use of modal verbs is mediated by the text's topical domain. The chapter then concludes with a summary of findings.

6.1 Finite and non-finite verb phrases

Verb phrases (VPs) can be finite or non-finite. Compared to finite VPs, which can vary for tense, aspect, modality, and voice, non-finite VPs are more compressed and implicit due to their being less grammatically flexible (Biber & Gray, 2016). Figure 6.1 displays the normalized rates of all VPs, finite VPs, and non-finite VPs across the RAs and SNAs.

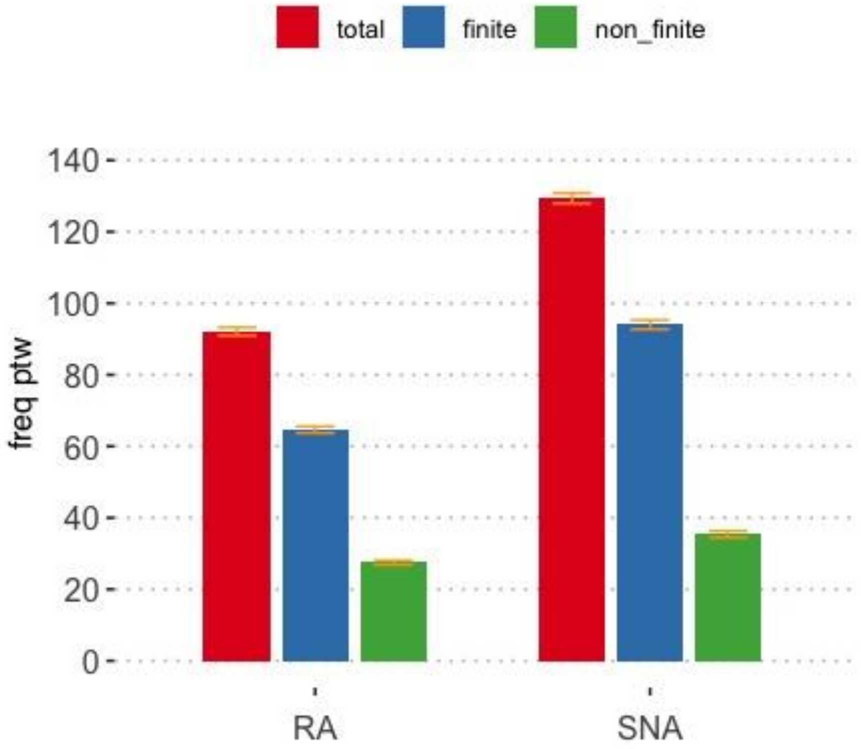


Figure 6.1 A bar graph illustrating the mean normalized frequencies of all VPs, finite VPs, and non-finite VPs across the corpora

Figure 6.1 shows that SNA writers used more VPs overall (129.37 ptw vs. 92.09 ptw, respectively), including more finite (94.01 ptw vs. 64.61 ptw) and non-finite (35.36 ptw vs. 27.48 ptw) VPs. The non-overlapping 95% CIs between corpora suggest that these differences are statistically significant. The difference in the overall rate of VPs is meaningful with a large effect size (*d* = 2.79).

Figure 6.2 provides a closer look at the non-finite VPs, revealing that *to* infinitive clauses and *-ing* clauses were used significantly more often by SNA writers, while *-ed* clauses were used marginally more frequently by RA writers.

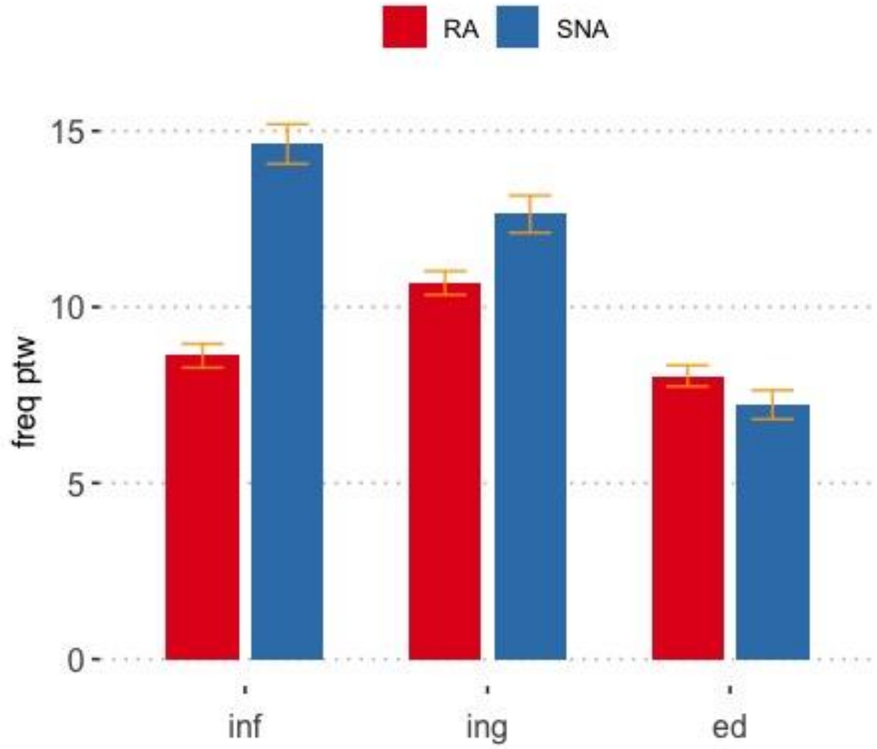


Figure 6.2 A bar graph illustrating mean normalized rates of three non-finite VP types across the registers

These statistics mean that a reader would, on average, come across 1 to 2 VPs for every 10 words in a given SNA. To illustrate what these numbers correspond to in actual texts, two excerpts from matching health articles are shown below. The SNA excerpt is shown first and then the RA (finite VPs **bolded**, non-finite VPs underlined):

Disarming a bacterial "secret weapon" could help fight superbugs

Australian scientists **have discovered** a previously unknown protective mechanism bacteria use to kill immune cells, opening up potential research pathways that **could**

> **help** doctors fight antibiotic-resistant superbugs. When the body **detects** a bacterial infection, one of its defense mechanisms **is** to release immune cells such as macrophages, which **hunt for** pathogens, then engulf and destroy them. But the bacteria **can fight back**, releasing a number of different toxins to target immune cells, many of which **are** well known and common targets for anti-superbug research - if we **can work out** how to block these toxins and disarm the bacteria’s defenses, we**'re** on the way to fighting antibiotic-resistant bugs.
>
> Health SNA #99

In this 107-word excerpt, the writer uses 20 diverse verb phrases, including tensed (e.g., *detects*), modalized (e.g., *could help*), and non-finite (e.g., *to target*) VPs. In contrast, the matching RA excerpt shows fewer and more uniform VPs:

> **Mitochondrial dysfunction caused by outer membrane vesicles from Gram-negative bacteria activates intrinsic apoptosis and inflammation**
>
> Outer membrane vesicles (OMVs) from Gram-negative bacteria **contribute to** pathogenesis, immune evasion, inflammation and immunity, but the mechanisms and roles of various OMV molecules **remain** incompletely understood. OMVs **trigger** inflammation as the lipid A moiety of lipopolysaccharide (LPS) **is sensed** by cell-surface Toll-like receptor 4 and cytosolic caspase-11. Active caspase-11 **liberates** the amino-terminal fragment of gasdermin D (GSDMD), which **causes** cell death (termed pyroptosis) and NOD-like receptor family and pyrin domain-containing protein 3 (NLRP3) inflammasome-mediated interleukin-1β (IL-1β) release via potassium (K+) efflux. However, OMVs from pathogenic bacteria **induce**

macrophage cell death and IL-1β secretion despite the expression of modified LPS that **evades** Toll-like receptor 4 and caspase-11 detection.

Health RA #99

This excerpt is also a bit over 100 words and also comes from the introduction. However, in contrast to the SNA excerpt, the RA one shows fewer than 10 VPs, most of which are finite and tensed. Instead of VPs are markers of literate, scientific writing in the natural sciences, including nominalizations and noun phrases with stacked modifiers. For example, the first independent clause from the above excerpt is a TV+PhO pattern (see Chapter 5), with nouns on either side modified by adjectives (*outer*), nouns (*OMVs*), and prepositional phrases (*from Gran-negative bacteria*). As a result, the discourse is difficult to parse with a non-expert eye. By contrast, the SNA excerpt uses several VPs in the form of complement, adverbial, and relative clauses. While still informational, these verbs help to guide non-experts in the understanding of what the researchers did in their research and how a certain biological process works.

Turning to types of non-finite verbs, the RA corpus showed a higher use of *-ed* clauses, a structure often used to modify nouns (Biber et al., 1999, p. 606). Because of the nominally heavy style of academic writing, these clauses also appear most often in academic writing (Biber & Gray, 2016, p. 92). While RA writers used a variety of verbs heading *-ed* clauses, most notably *used* (.24 ptw) and *associated* (.23 ptw), SNA writers relied almost exclusively on *publish* (.74 ptw) and *call* (.54 ptw), which serve unique functions in SNA writing. By contrast, *to* infinitive clauses and *-ing* clauses are frequently used as either noun or verb complements (Biber et al., 1999, Ch. 9), and thus were more common in SNAs, where verbs were particularly frequent (see Figure 6.1).

This distribution of non-finite VPs across the corpora roughly corresponds to what Biber and Gray (2016) refer to as the phrasal versus clausal continuum. On this continuum, academic writing tends to prefer phrasal features embedded within noun phrases, while spoken conversation tends to prefer finite dependent clauses as clause elements. The slight preference for *-ed* clauses, which frequently function as noun modifiers, by RA writers and *to* infinitives and *-ing* clauses, which more freely occur as clause elements, by SNA writers can be approximately mapped onto this continuum, where SNAs lean toward the end with conversational features and RAs toward the end with literate, academic features.

6.2 Variation in Tense and Aspect

Turning now to variation in finite verb phrases, this section examines variation in tense and aspect. Figure 6.3 displays the normalized frequencies of past and present tense verbs (regardless of aspect), as well as perfect and progressive aspect verbs (regardless of tense).

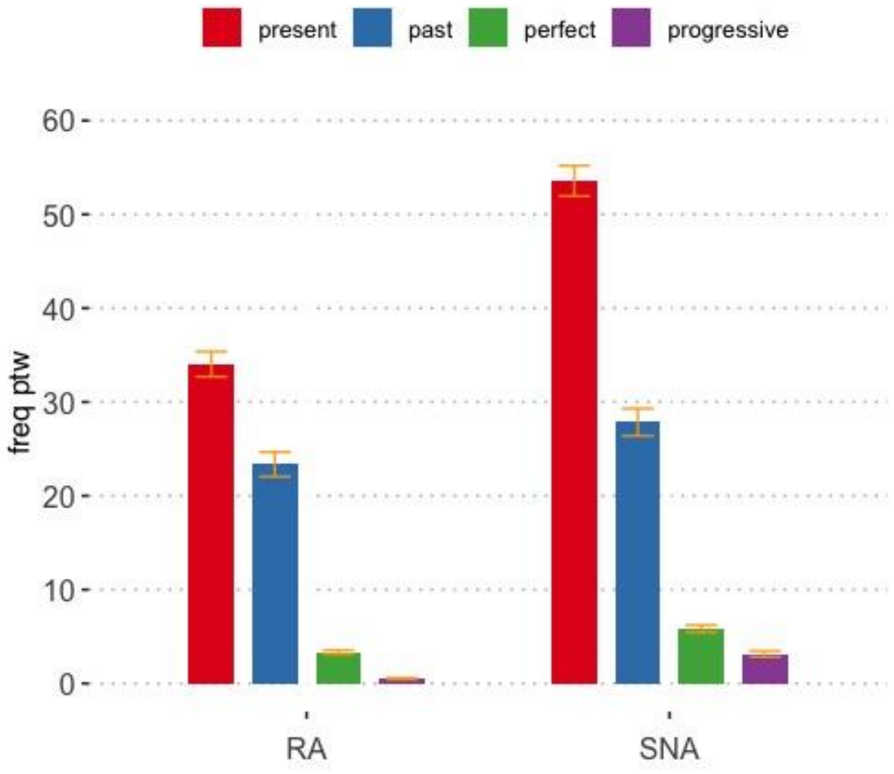


Figure 6.3 A bar graph depicting rates of all present tense, all past tense, all perfect aspect, and all progressive aspect verbs across the registers

Figure 6.3 reveals that the SNA corpus used more present tense, past tense, perfect aspect, and progressive aspect verbs. The non-overlapping 95% CIs suggest that these differences are statistically significant. These four grammatical contexts were also significantly associated with register ($\chi 2(3) = 687.11$, $p < .001$, Cramer's $V = .11$). The most important association was between the progressive aspect and SNA corpus (adjusted residual = 20.19), followed by the past tense and RA corpus (adjusted residual = 18.59), present tense and SNA corpus (adjusted residual = 9.24), and perfect aspect and SNA corpus (adjusted residual = 5.07). Below, these findings are explored in more detail.

While SNA writers used more verbs overall, their articles were most associated with verbal markers of non-past time, such as the present tense and progressive aspect. To explore these associations further, the ten most frequent present tense verbs in both registers are presented in Table 6.1.

Table 6.1 The ten most frequent present tense verbs across the registers

RA corpus		SNA corpus	
Verb	Freq ptw (raw)	Verb	Freq ptw (raw)
be	9 (18109)	be	14.65 (4122)
show	1.13 (2269)	**say**	2.92 (821)
have	0.97 (1952)	have	1.79 (505)
suggest	0.65 (1303)	suggest	0.94 (264)
use	0.48 (961)	**know**	0.77 (218)
provide	0.4 (805)	**make**	0.77 (218)
find	0.37 (755)	show	0.74 (209)
indicate	0.36 (735)	**need**	0.65 (182)
include	0.33 (661)	**think**	0.64 (179)
increase	0.28 (560)	find	0.54 (152)

Note: **bolded** words are those that are unique to their respective column

Half of the verbs in Table 6.1 are unique to their respective columns. That is, they were among the most frequently used verbs in one register but not the other. Additionally, these verbs are diverse in semantic domain, including verbs expressing research findings (*find, show*), conclusions (*indicate*, *suggest*), methods (*use*), mental activities (*know*, *think*), communicative acts (*say*), interpretation (*provide*), explanation (*have, include, increase, make*), and characterization/identification (*be*).

Present tense verbs expressing mental (*know, think*) and verbal (*say*) acts were more characteristic of the SNA corpus, reiterating the tendency of SNA writers to use language that reports the words and ideas of others (see section 5.2.3). These verbs often had humans as their grammatical subject and could be found within interview excerpts, as shown in excerpt (6.1) (present tense verbs **bolded**, human subjects underlined):

(6.1) "Usually, when we **think** about food, we **think** about the taste and the aroma and, of course, the sight of it," Danni Peng-Li, a doctoral student in the department of food science at Aarhus University and the study's lead author, told The Academic Times. [Psychology SNA #25]

Informants connected the use of present tense with the function of news being to report on recent events. One informant put it this way:

> *News media in general wants the reader to think that this is urgent and current, so that's why you end up seeing more present tense […] News is very present*-ist. *You read the story today, and then you forget about it tomorrow. A lot of the value of news is immediate.* [Informant #4]

In other words, science news attracts readers by highlighting the urgent and current nature of recent research, a feature which is constructed by and reflected in the higher frequency of present tense verbs.

Common present tense verbs not referring to mental or verbal acts include the copular *be*, which occurred nearly 6 more times per 1000 words in SNAs than RAs. This finding is insightful relative to the analysis of verb patterns in Chapter 5. There, both corpora were shown to use similar overall rates of copular verb patterns. Thus, the greater use of *be* here and elsewhere in this chapter section likely reflects the fact that, while both registers used similar amounts of *be* as main verbs of independent clauses, SNAs also often used *be* in dependent clauses, which were not included in Chapter 5's analysis.

Among other purposes, *be* often helped define terms (6.2) and express personal stance (6.3) (present tense *be* verbs **bolded**):

(6.2) Metabolic proteins **are** essentially chains of amino acids that convert nutrients into energy in any living organism [Space SNA #251].

(6.3) To understand why this new mass information **is** so crucial, we can look to our own galaxy, the Milky Way. [Space SNA #267]

Arguably, both (6.2) and (6.3) function to guide the reader in understanding the content of the article. In (6.2), the copular *be* defines an important, though not necessarily widely known, definition of a key term. The addition of *essentially* also highlights the writer's intention to make the content as digestible to the reader as possible. In (6.3), the writer packages research findings into a nominalization (*this new mass information*) and uses *is* to link that finding to an interpretation of its significance—that it is *so crucial*. Again, the writer appears mindful of the reader, not just summarizing an RA but highlighting key information by labelling it as important.

Like the present tense, the higher rates of perfect and progressive aspect verbs in SNAs also suggest an important role for reference to the present time in this register. The (present) perfect aspect represents past-time-related-to-the-present-time (Leech, 2004). Thus, while it incorporates a period preceding the present, it also highlights a resultant state in the present. The (present) progressive aspect views an event as durative and often in progress at the time of utterance (Leech, 2004), and is thus associated with present time reference.

To better understand the uses of these features in texts, Table 6.2 presents the ten most frequent present perfect and present progressive verbs in the registers.

Table 6.2 The ten most frequent present perfect and present progressive verbs across the registers

RA corpus				SNA corpus			
Present perfect		Present progressive		Present perfect		Present progressive	
Verb	Freq ptw (raw)	Verb	Freq ptw (raw)	Verb	Freq ptw (raw)	Verb	Freq ptw (raw)
show	0.23 (467)	**increase**	0.01 (20)	be	0.34 (95)	**look**	0.07 (21)
be	0.13 (265)	**miss**	0.01 (17)	find	0.28 (79)	**happen**	0.07 (19)
find	0.11 (229)	**lack**	0.01 (14)	show	0.22 (63)	**work**	0.06 (18)
report	0.08 (153)	**become**	0.01 (13)	**publish**	0.13 (37)	**get**	0.06 (17)
use	0.07 (135)	**drive**	0.01 (13)	**see**	0.1 (28)	go	0.06 (16)
demonstrate	0.06 (126)	**experience**	0.01 (12)	**develop**	0.08 (23)	**come**	0.05 (15)
observe	0.05 (94)	**use**	0.01 (12)	use	0.07 (20)	**start**	0.05 (15)
focus	0.04 (90)	do	0.00 (9)	**link**	0.07 (19)	**see**	0.05(13)
propose	0.04 (87)	**emerge**	0.00 (9)	**become**	0.06 (18)	**try**	0.05 (13)
suggest	0.04 (86)	look	0.00 (9)	**discover**	0.06 (18)	do	0.04 (10)

Note: **bolded** words are those that are unique to their respective column

Looking first at present perfect verbs, most of these verbs were unique to their respective columns, while others (e.g., *show, find, use*) were found in both. In terms of their purposes, RA writers often used these verbs to reference previous literature, as (6.4) demonstrates (present perfect verb **bolded**, previous literature underlined):

(6.4) In fact, only recently, contextual and multisensory effects on consumer behavior and food choice **have been** systematically **investigated** (Krishna & Schwarz, 2014; Spence, 2012). [Psychology RA #25]

This excerpt describes a situation in which past research has changed the current understanding of consumer food choice habits. A similar function can also be seen in SNAs, where the resultant state often equates to ‘we now know something new’ (6.5) or ‘we have created something new’ (6.6) (present perfect verb **bolded**):

(6.5) Decades of watching these space rocks plunge into or graze the Sun **have allowed** researchers to suss out the chemicals within comets. [Space SNA #156]

(6.6) And that’s just what a team of researchers **has developed**, in the form of novel nanocrystals that allow the radiation dose from a diagnostic X-ray to be much lower, while also enabling higher resolution images at a lower cost. [Health SNA #180]

Thus, the present perfect seems to function similarly in both registers, though these sentences are more frequent overall in the SNA corpus. Their greater use, especially relative to the simple past, is likely due to the writers’ desire to build proximity and relevance. For example, although the researchers in (6.5) completed “sussing out” the chemical composition of comets prior to the writing of the SNA, it is the resultant state of having this knowledge that is important—not the act of identifying the chemicals itself. Thus, the perfect aspect creates a stronger link between a past situation and the present time.

The progressive aspect was rare overall but most frequent in SNAs, due in part to the higher rate of quoted speech. In these quotes, researchers often highlight a developing trend that affects general populations. As a result, the article feels more timely, since it reports on current thinking and phenomena, and more relevant, since the topic is made relevant to groups outside of those who research it. Consider the following extract (present progressive verbs **bolded**):

(6.7) “I think awareness **is growing** across many disciplines, as well as in society more broadly, that mental and physical health are connected - or that, in other words, mental health is health,” Ditmars said. [Psychology SNA #97]

This extract reflects many of the features of SNA writing discussed thus far. The writer calls on the expertise of a researcher in the article (*Ditmars*), directly quotes them using a FCl+S|TV verb pattern (*"…," Ditmars said*), reports the mental state of that researcher (*I think*), and shows a present progressive verb reporting a developing trend that affects broader society.

In some cases, the present progressive reflects the fact that SNA writers are often interested in the next steps that the researchers will take and thus look beyond the results of the immediate study (6.8) (present progressive verbs **bolded**):

(6.8) All of this, of course, needs to be proven in human subjects, but the researchers **are** already **working towards** several human clinical trials in both adults and infants. [Health SNA #213]

The progressive verb phrase in (6.8) describes an activity that the researchers are engaged in but have yet to complete, reflecting another way in which SNAs can differ from RAs. That is, SNAs are more interested in next steps and applications than RAs, which tend to be more confined to describing the results of the present research.

By contrast, RAs more often used past tense verbs, in large part due to the greater presence of methodological description. Table 6.3 displays the ten most frequent past tense verbs in both the registers.

Table 6.3 The ten most frequent past tense verbs across the registers

RA corpus		SNA corpus	
Verb	Freq ptw (raw)	Verb	Freq ptw (raw)
be	3.31 (6677)	be	3.73 (1050)
use	0.85 (1719)	**say**	1.88 (529)
have	0.65 (1314)	find	1.46 (412)

find	0.65 (1310)	have	0.7 (198)
show	0.48 (963)	**publish**	0.55 (154)
include	0.36 (716)	show	0.48 (135)
perform	0.35 (702)	report	0.38 (107)
observe	0.34 (702)	use	0.37 (105)
report	0.34 (683)	**tell**	0.32 (89)
measure	0.23 (454)	**author**	0.3 (85)

Note: **bolded** words are those that are unique to their respective column

Previous research has shown that the past tense often functions to report methods, results, and past literature in RA writing (Heslot, 1982; Swales & Feak, 2012), and the same was true for the RAs of this dissertation. Table 6.3 shows that several verbs used to describe methodology, such as *use*, *perform,* and *measure*, were frequent in the RA corpus but less frequent in or absent from the SNA column. As already shown in section 5.2.1, informants also noted that methodological description is often minimal in SNA writing. Similarly, another informant noted that, "[The literature review] doesn't exist" in SNAs (Informant #1), which might explain the more limited role of the past tense in SNAs.

By contrast, the SNA column in Table 6.3 shows several verbs describing acts of communication, including *say, report, tell,* and *author*. *Say* was notably frequent, occurring about ten times more often in the SNA corpus than the RA corpus, while *tell* was a frequent alternative when also including the indirect object of the verb pattern. While *say* and *tell* usually communicated direct or indirect speech, *report* identified study findings, especially those pertaining to humans in experimental conditions. It is perhaps less surprising, then, that this verb was also common in the RA corpus (see Table 6.3). Lastly, *author* frequently occurred in the passive voice, serving to introduce the authors of the study being reviewed (6.9):

(6.9) The study, “Exposure to Televised Political Campaign Advertisements Aired in the United States 2015-2016 Election Cycle and Psychological Distress,” published April 3 in Social Science & Medicine, **was authored** by Jeff Niederdeppe, Rosemary J. Avery, Jiawei Liu, Brendan Welch […] [Psychology SNA #196]

In short, RAs used more past tense verbs, which were frequently used to contextualize their study in past literature and detail study methodology. By contrast, SNAs used past tense verbs less frequently and more often for the purpose of reporting others’ words, while reporting study findings with verbs like *reported* and *showed* were similarly common in both corpora.

6.3 Variation in grammatical voice

Many verb phrases can be grammatically active or grammatically passive, the choice being conditioned by their contexts of use and the intentions of the writer. ‘Short’ passives are those which exclude a *by* phrase containing the agent of the verb, while ‘long’ passives retain the *by* phrase. Figure 6.4 illustrates the mean normalized frequencies of these passive verbs across the registers.

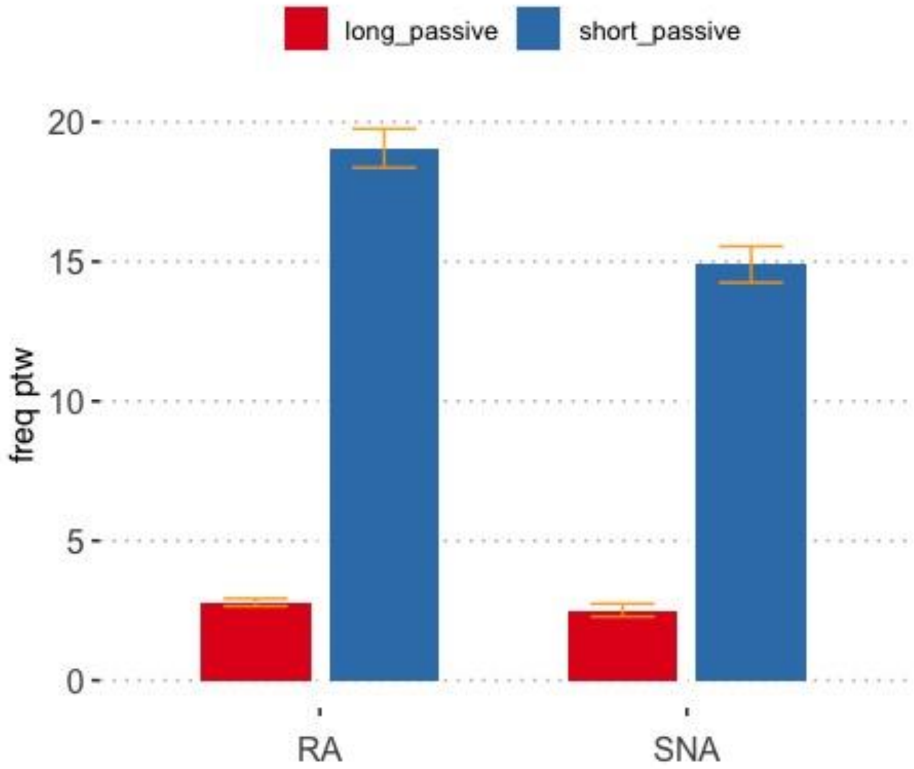


Figure 6.4 A bar graph illustrating mean normalized rates of long passive and short passive verb phrases across the registers

Figure 6.4 shows that the RA corpus used more short passive VPs and more long passive VPs, while the 95% CIs suggest that these differences are significant with respect to short passives. Moreover, these frequency differences are meaningful with a medium effect size (*d* = 0.61). Both SNA and RA writers used similar rates of passive verbs in the present tense (5.73 ptw vs. 6.24 ptw) and simple aspect (1.38 ptw vs. 1.34 ptw), with the most significant difference being the RAs' use of past tense passive verbs (3.84 ptw vs. 7.71 ptw). In short, RAs used more passive voice verbs overall, with short past tense passive verbs being particularly descriptive of this register.

Informants expressed strong beliefs about the use of grammatical voice in science news writing. Several identified it as a core stylistic concern of SNAs. Consider the following excerpts from two informants:

> *[The use of the passive voice] is like Science Writing 101. It's always like all these research articles are in passive voice, and it's our job as science journalists to turn them all into active voice, because I do think that it generally is more effective for pulling readers in* [Informant #5]

> *It's not considered good news style to use the passive voice. You're supposed to avoid it. That's something you would learn on day one in journalism school.* [Informant #4]

These excerpts suggest that some science journalists are taught to be concerned about the use of grammatical voice in science writing. Active voice is emphasized in part because it creates a more engaging style of writing, while passive voice reflects a traditional, impersonal style of academic discourse.

However, Figure 6.4 also shows that the passive voice is still prevalent in SNAs. For example, the difference in use of overall verb phrases is about three times more meaningful than the difference in the use of passives (see section 6.1). While there are likely several reasons for the use of passive verbs in SNAs, one informant described a benefit of using the passive voice this way:

> *Basically,* [a scientist] *needs to write things in a way to convince everyone, "I'm a legit scientist," and passive voice is a big part of that. [...] When you translate that into science news, there's conflict there because, on one hand, active voice is the thing that people find engaging and exciting. You write active to keep it dynamic and to keep the article moving. But you also on the other hand want it to still feel scientific because, if it*

> *feels too active and too dynamic, then—it's that whole pseudoscience thing—it just doesn't feel scientific.* [Informant #2]

This conflict between sounding engaging and sounding scientific may help to explain why passive voice is seen by some as antithetical to good science news writing while also being somewhat common in discourse.

Another explanation may be related to *which verbs* are used in the passive voice. That is, some verb-voice combinations may be seen by SNA writers as more 'scientific' while others regularly appear in their passive form without concern. To explore this hypothesis further, see Table 6.4 below, which presents the most frequent (finite) passive voice verbs. Note, verbs are also reported alongside any statistical associations (based on a Chi-square test of independence) with different grammatical contexts to add additional insight.

Table 6.4 The ten most frequent finite passive voice verbs across the registers

RA corpus		SNA corpus	
Verb	Freq ptw (raw)	Verb	Freq ptw (raw)
use^{Pst}	0.69 (1389)	**publish**^{Pst}	0.67 (193)
associate^{Pr}	0.45 (902)	find^{Pst}	0.36 (102)
observe^{Pst}	0.32 (648)	use^{Md}	0.29 (82)
find^{Md}	0.31 (630)	**author**^{Pst}	0.29 (81)
show^{Pr}	0.3 (594)	associate	0.24 (67)
perform^{Pst}	0.26 (529)	**make**^{Pr}	0.21 (60)
consider^{Md}	0.18 (369)	**involve**^{Pst}	0.2 (57)
measure^{Pst}	0.18 (359)	**know**^{Pr}	0.18 (51)
calculate^{Pst}	0.18 (357)	**link**	0.15 (43)
expect^{Pr}	0.18 (352)	**think**	0.14 (40)

Note: **bolded** words are those that are unique to their respective column
^{Pr} = positively associated with present tense
^{Pst} = positively associated with past tense
^{Md} = positively associated with modalized verb phrase

Table 6.4 shows that many of the most frequent passive voice verbs are unique to their respective columns. Moreover, several of the verbs shared across the columns, such as *use, associate,* and *find*, show different associations with their preferred grammatical contexts. These findings support the above hypothesis that the registers use the passive voice differently in quantitative and qualitative ways. Below, two domains of verbs, namely activity and mental verbs, are explored in more detail.

Several activity verbs were associated with the past tense, where they served to describe study methodology. These verbs were most frequent in the RA corpus, including *use, observe, perform, measure,* and *calculate*, while *use* and *involve* were among the most common verbs in the SNA corpus. Excerpt (6.10) illustrates the use of past tense passive verbs describing research methodology in an RA (past passive verbs **bolded**):

(6.10) A thin-ruled grid **was adhered to** the bottom of the plate to facilitate precise cuts between the zones (Fig. 1). Plants **were kept** submerged in RNAlater at all times. Three independent seedlings **were used as** three biological replicates for each of the transcriptome analyses. All seedlings **were grown** on the same plate. [Space RA #117]

This excerpt shows four passive verb phrases serving to recount methodological procedure. The assumed agent of the verbs is the research group and thus may be omitted. Moreover, by adopting the passive voice, the writers alter the message of the sentences, making the research objects (*grid, plants, seedlings*) the ‘theme’ of the message rather than the researchers (see Halliday & Matthiessen, 2004, Ch. 3).

It follows that these verbs are the sort that SNA writers may want to avoid due to their sounding particularly ‘scientific’. SNAs desire to tell stories about the scientists involved in the highlighted research, so by adopting the passive voice with activity verbs controlling human agents, removing those agents from the context betrays one of the main purposes of the register.

Additional evidence can be seen in the use of *use* by SNA writers. This verb was associated with the past tense in RAs but modal verbs in SNAs. While the past passive form of *use* often describes the methods adopted in a particular study, modal verb + *use* instead emphasizes the implications of a new device, model, or some other information in unreal scenarios. Consider the following excerpt from an SNA (modal verb + passive *use* **bolded**):

(6.11) The researchers say that future studies will incorporate this metric into the experiment. If it all works out, the process **could be used** for human implants within five to 10 years. [Health SNA #66]

Could locates the message of (6.11) not in the past but rather in an unreal scenario. This scenario describes how the findings could be applied contingent upon further research being completed. The implied agents of modal verb + *use* are diverse, including greater humanity or specific academic fields, but are rarely the individual research teams themselves.

Like activity verbs, several mental verbs were found in both registers, including *find, know,* and *think* in the SNA corpus, and *find, observe, consider,* and *expect* in the RA corpus. Notably, only *find* is shared across the columns in Table 6.4. *Know* and *think* are particularly revealing of differences in the use of passive voice between the registers. These verbs often identify commonly held knowledge in academic domains, with the commonness of the knowledge being a motivating factor for the use of passive voice. Consider the following excerpt (passive voice mental verbs **bolded**):

(6.12) Given the SARS-CoV-2 virus behind the outbreak **is** largely **thought** to be a zoonotic pathogen that spread to humans from animals, the postponement is just another example of how pressing these conservation issues are, the team says. [Health SNA #327]

Here, the writer presents the information that SARS-CoV-2 is a zoonotic pathogen as *given* knowledge *largely thought* by a vague but decidedly informed community. Sentences like this one subtly and succinctly inform readers about the state of knowledge in an academic field. Sentences with the passive form of *observe* or *expect* in RAs can serve a similar purpose, but the information following the verb is less generally accepted, more esoteric, and often attributed to specific studies via citation. In some cases, these verbs did serve a different purpose, including providing rationale for a methodological choice (6.13).

(6.13) We used a set of six primers designed to target the highly conserved recF gene [44, 57]. These primers are specific to Salmonella sp., and thus **are** not **expected** to hybridize to DNA of unrelated pathogens. [Health RA #29]

The RA writer in (6.13) uses the passive *are expected* in the methods section of their article, and, unlike in previous extracts, the assumed agent is arguably the writers themselves. Another difference was in the grammatical associations of *consider* and *find*. These verbs were associated with modal verbs in the RA corpus, while *find* was associated with the past tense in the SNA corpus and *consider* was not frequent enough to appear in the SNA column. For either verb, the passive clause often had the article's reader as the assumed subject:

(6.14) The following limitations **should be considered** when interpreting the findings of this study. [Psychology RA #107]

(6.15) A glossary of key terms **can be found** in Table 1. [Space RA #143]

The likely intended agent of (6.14) is the reader, as only readers of that RA could consider the limitations provided in the article. The same logic applies to (6.15), where the passive verb points readers to information not included in the prose of the article.

In short, finite passive verbs were more frequent in RAs than SNAs. SNA writers hold similar opinions about the effect of grammatical voice on their writing, namely that more passive verbs equate to more scientific, less enjoyable reading, while active voice verbs are better for writing stories in with a reader-friendly style. Although SNAs still use a not insignificant amount of passive verbs, qualitative differences between the types of verbs used in the passive form and their purposes in discourse differ.

6.4 Variation in modal verbs

Some finite verb phrases have modal verbs and thus are not marked for tense but rather modality. This section investigates the frequencies and uses of finite verb phrases with central or semi-modal verbs across the registers. Figure 6.5 illustrates the mean normalized frequencies of eight central modal verbs across the registers.

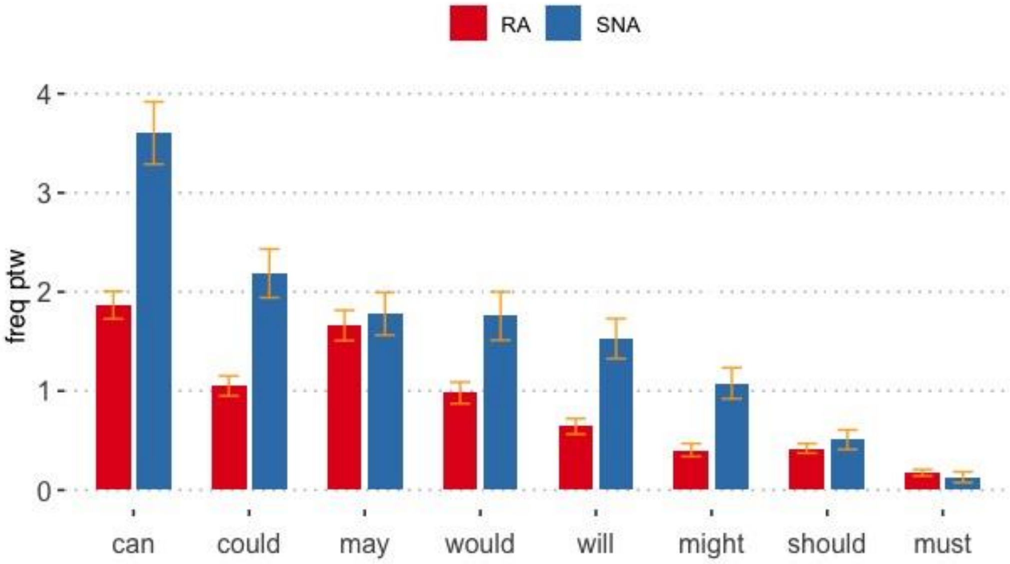


Figure 6.5 A bar graph illustrating the normalized frequencies of eight modal verbs across registers

The non-overlapping 95% CIs in Figure 6.5 suggest significant differences in the rates of *can, could, would, will,* and *might* between the registers such that SNA writers use more of them. The greater use of *can* in SNAs was particularly meaningful, while *may, should,* and *must* were used with similar frequency by both RA and SNA writers.

Three semi-modal verbs were also investigated. Figure 6.6 illustrates the rates of these semi-modal verbs across the registers.

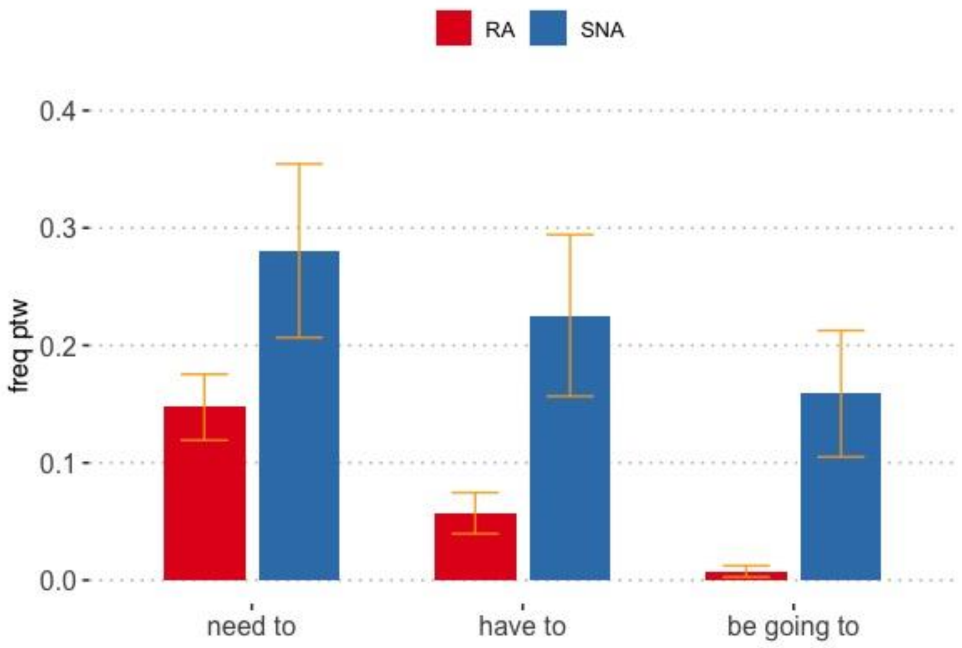


Figure 6.6 A bar graph illustrating the mean normalized rates of three semi-modal verbs across the registers

The non-overlapping 95% CIs in Figure 6.6 for *need to, have to*, and *be going to* suggest significant differences such that SNAs use more of these features. Together, the greater use of central and semi-modal verbs in the SNA corpus was meaningful with a large effect size (*d* = 1.08).

Informants ascribed the greater use of modal verbs to two characteristics of science news writing. First, they described wanting to accurately portray science as incremental and inconclusive. Consider the following excerpt from one informant:

> *Thinking as a reporter, I think* [modal verbs] *are necessary to express to the audience that this is not definite, that what the study found, while relevant and interesting, it's not necessarily conclusive, which I mean, science overall is not conclusive. But the audience might not be aware of this, so it's always good to express that.* [Informant #3]

Several informants stated a position like this one, namely that individual studies are rarely conclusive and that science overall is always changing with new insights into how the world works. For SNA writers, this attitude is communicated through modal verbs such as *may* or *can*.

Another characteristic affecting the use of modal verbs in SNAs is reader background knowledge. That is, because the reader’s experience with scientific research is unknown to the writer, writers may use more modal verbs to avoid sounding overly confident in their claims. As one informant put it, “I think I trust my audience a lot less when I'm writing for popular science and that would affect me using a lot of [modal verbs]” (Informant #6). Another informant put it this way:

> *Research articles have the luxury of explaining the statistics of it. And that self explains, “Well, there’s this sort of chance,” or “there’s a hazard ratio that says this,”* […] *so everything’s going to be translated from those sorts of hazard ratios, or sigmas, or a* p *value, and you need to take that number that makes sense to scientists and kind of word it, and that’s probably where you’ll probably get a lot of* could*’s,* may*’s,* maybe*’s.*
> [Informant #2]

These excerpts demonstrate audience awareness on the part of SNA writers. Some writers may feel added pressure to hedge their claims and interpretations of studies so as to not be misconstrued by their readers. Additionally, RAs are generally long, details texts written for technical audiences, allowing them to express tentativeness through semiotic resources including but not limited to statistics. Due to the technical knowledge required for interpreting some statistics, SNAs tend not to rely on this resource.

Looking more closely at the distribution of modal verbs across the registers, two semantic categories stand out as being particularly frequent in SNAs. First, the modals *can, could, may,* and *might*, which express permission/possibility/ability (hereafter, ‘possibility modals’), occurred 8.86 ptw in the SNA corpus but 5.04 ptw in the RA corpus. *Can* was particularly frequent, one reason for which may be because *can* lends itself to the rhetorical appeal of application. Appeal to application helps science communicators demonstrate the value of science through emphasizing how it can be applied to practical problems (Fahnestock, 1986). Consider the following excerpts from two SNAs (*can* **bolded**):

(6.16) Now, a team of researchers in Japan have proposed a new, physics-based technique that **can** predict, in many cases, whether or not a solar flare is about to occur. Not only that, the technique **can** also pinpoint where the flare will erupt on the Sun, as well as roughly how powerful it will be. [Space SNA #58]

(6.17) Even today in hospitals, doctors type in some of your data, and the algorithm **can** tell you’re at higher risk for heart disease or certain types of cancer. [Psychology SNA #106]

The bolded verbs in (6.16) and (6.17) highlight the application of a recent breakthrough or development, such as a simulation technique or algorithm. The appeal is that such applications could have a more-or-less direct impact on the average reader’s life. Even in excerpt (6.16), which details a computational technique for predicting solar flares, the writer quickly addresses how the finding could affect the reader. In the following sentence, the writer says, it may help “protect us from the next Carrington Event,” referencing a disruptive geomagnetic storm that occurred in 1859.

Ostensibly, *could* could also be used for this function. However, in the SNA corpus, this modal verb infrequently paired with subjects lending themselves to discussion of applications, like those in (6.16) and (6.17). Rather, the subjects of verb phrases with *could* were not often models, techniques, or other research objects. Moreover, when the sentences could be understood as an appeal to application, the more tentative nature of *could* made it less persuasive, as illustrated in (6.18):

(6.18) Besides, repairs are probably not going to cut it at this point, although they **could** buy us precious time to offload the oil. [Environment SNA #316]

The scenario described by the verb phrase with *could* in excerpt (6.18) is hypothetical, contingent upon some other condition being met in order to take effect. Thus, it is less forceful as an appeal to application.

Other possibility modals, such as *may* and *might*, were also frequent relative to other semantic classes of modal verbs. One apparent use of these verbs is to involve the reader intellectually with the article. *May* and *might* mark propositions as hypotheses (De Waard & Maat, 2012), to be presented to the reader so that they use information in the article and their background knowledge to come to their own conclusions. Consider the following excerpt from the beginning of a health SNA (possibility modals **bolded**):

(6.19) Spewing virus-laden droplets **may** not be the only way animals **can** spread some viruses through the air. Viruses like influenza **might** also hitch a ride on dust and other microscopic particles, a study in guinea pigs suggests. [Health SNA #7]

In the first sentence of (6.19), the writer anticipates the reader's prior knowledge of virus transmission and constructs the expectation that the article may provide evidence to the contrary. The reader will then need to read the remainder of the article, consider the evidence, and come to a conclusion whether animals do indeed spread air-borne viruses via additional modes. Thus, while these modals serve as hedging devices to protect the writer's interpretation of the study, they can also serve to intellectually involve the reader in the article's content.

The second and smaller semantic class of modal verbs to be examined here is the prediction modals, comprised of *will, would,* and *be going to,* which were more than twice as frequent in the SNA corpus than in the RA corpus. The use of these verbs to make logical predictions, as opposed to mark personal volition, is particularly common in academic writing (Biber et al., 1999, p. 496), suggesting an important role for this function in SNAs. In either register, *would* largely served to speculate on hypothetical scenarios and *will* on future scenarios. Consider the following excerpts (prediction modals **bolded**):

(6.20) One limitation of the model, however, is that it doesn't reproduce non-active regions of the Sun with weak magnetic fields, which Ulrich says **would** be necessary to get a comprehensive picture of the Sun's magnetic field at any given moment, past or present. Though he admits it might not be possible to solve, he adds "it certainly **would** be worth checking out." [Space SNA #119]

(6.21) "Hopefully this means that it **will** be more effective at reducing poverty, providing stability and improving child and family health, and we'**ll** have to make sure to do studies to ensure it's working as it should." [Psychology SNA #234]

Sentences like those in (6.20) and (6.21) frequently have two characteristics in common. First, they are placed near the ends of articles, where they are used to consider applications and implications of study findings. For example, the second instance of *would* in (6.20) indicates an area of future research that would improve the study's model, and the modal verb *will* in (6.21) expresses desired future outcomes of a tax credit program and how future research could investigate its effects. Second, the predictions proposed by these verbs tended to be made by researchers rather than SNA writers. For example, the bolded verbs in (6.20) and (6.21) are found in direct and indirect reports of the researchers' words.

In short, SNA writers use many modal verbs in their articles. Modals expressing tentativeness are often used to hedge interpretations and invite reader involvement with the content of the article. More confident predictions are usually attributed to researchers rather than SNA writers, maintaining SNA writers' tendency to avoid misinterpreting or sensationalizing science.

6.5 Cluster analysis

As in Chapter 5, I performed a cluster analysis to group corpus texts together based on their use of features of short verb phrase variation. I grouped the possible variations into six variables, namely finite present tense verbs in active voice, finite past tense verbs in active voice, finite non-simple aspect verbs in active voice, finite modal verbs in active voice, non-finite verbs in active voice, and passive voice verbs. These variables fairly neatly capture variation in verb phrase finiteness, tense, aspect, voice, and modality. All 800 texts (400 RAs and 400 SNAs) were included in the analysis.

6.5.1 A two-cluster solution

A dendrogram was produced and visually inspected to determine the optimal number of clusters (Figure 6.7).

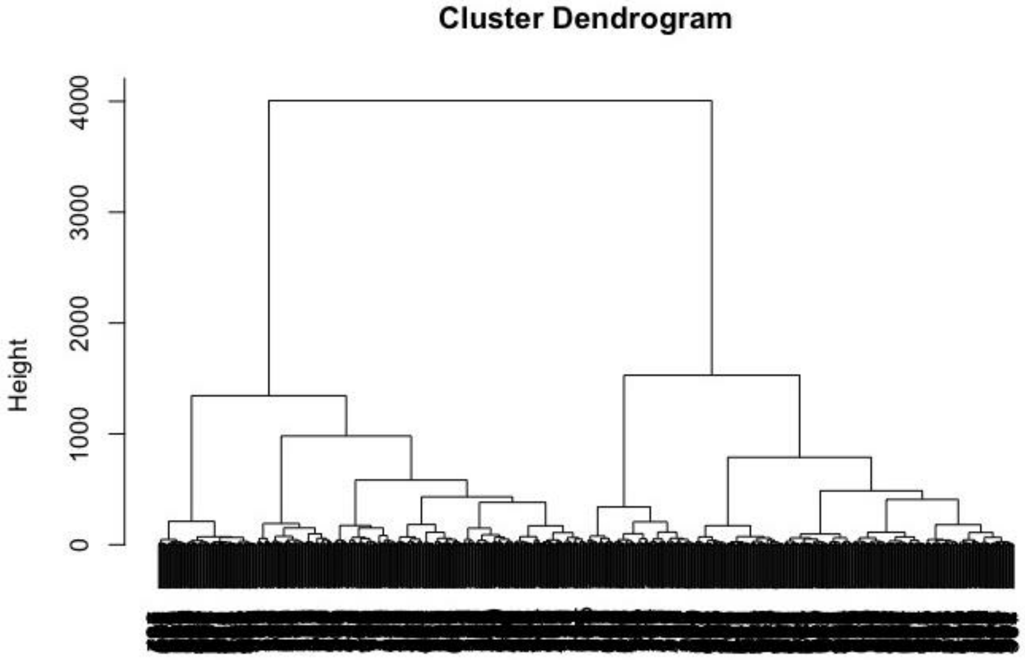


Figure 6.7 A dendrogram illustrating the clustering of 800 RA and SNA texts into groups at different levels

The dendrogram in Figure 6.7 suggests that a two-cluster solution is the most optimal one for grouping the 800 texts, since a horizontal line drawn across unit ~2500 of the y-axis neatly classifies all texts into two clusters. Figure 6.8 shows the proportion of each register in each cluster.

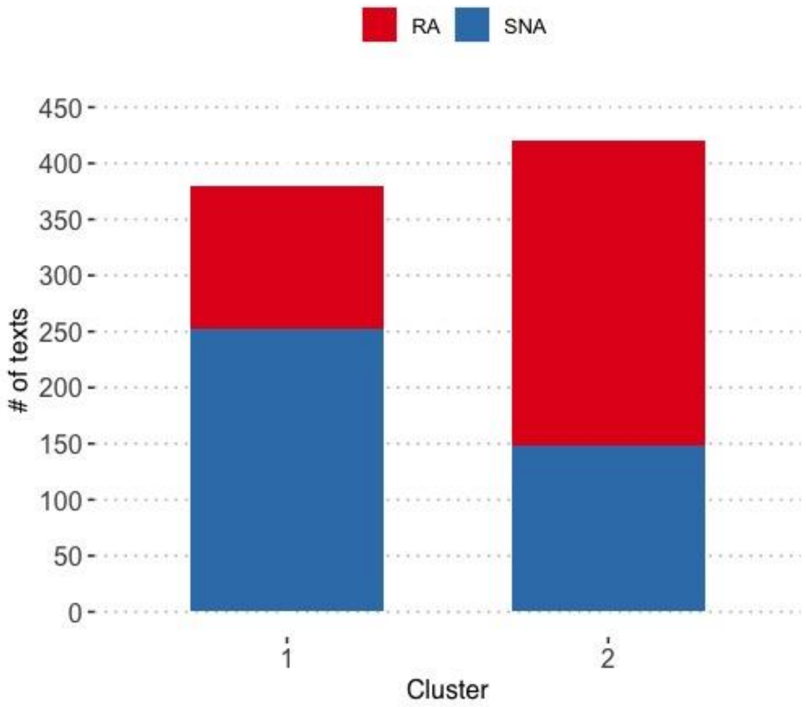


Figure 6.8 A bar graph showing the number of texts from each register by cluster

According to Figure 6.8, Cluster 1 is comprised mainly of SNAs and Cluster 2 mainly of RAs; however, each cluster also has a number of texts from the other register, revealing that the clusters do not simply split on register.

Finally, Figure 6.9 displays the proportions of the four topical categories in each cluster.

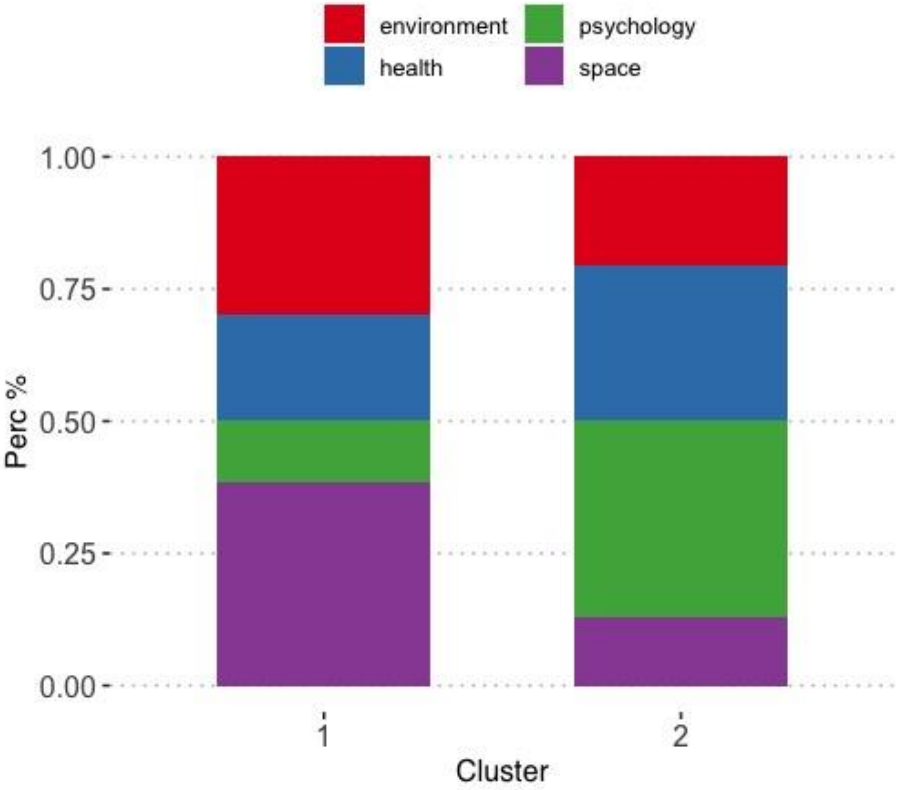


Figure 6.9 A proportional bar graph illustrating the percentages of topical category by cluster

Figure 6.9 reveals different profiles of topical domain across the clusters. The most distinct difference is the higher proportion of space texts in Cluster 1 and psychology texts in Cluster 2, while a more moderate difference is the higher proportion of environment texts in Cluster 1 and health texts of Cluster 2.

Next, in section 6.2.2, I explore the meaning of the two clusters by detailing the features that describe them and providing longer excerpts to illustrate their discourse.

6.5.2 Explaining the two clusters

To understand how verb phrase variation relates to the two clusters, mean normalized frequencies for the six variables were calculated for each cluster. Each cluster was then defined by the features that were most frequent in that cluster (Figure 6.10).

Cluster 1	Finite present tense verbs in active voice*, finite non-simple aspect verbs in active voice*, finite modal verbs in active voice*, non-finite verbs in active voice
Cluster 2	Finite past tense verbs in active voice*, all passive voice verbs*

*Significantly more frequent in this cluster relative to the other cluster, based on 95% CIs

Figure 6.10 The most frequent features of short verb phrase variation found in each cluster

Figure 6.10 shows that Cluster 1 is defined by more present tense verbs, verbs with non-simple aspect, modal verbs, and non-finite verbs, all in the active voice. By contrast, Cluster 2 is defined by significantly more past tense verbs and passive voice verbs. These profiles broadly reflect the findings of previous sections of this chapter. Namely, SNAs use more present tense, non-simple aspect, and modal verb phrases than RAs, which rely on past tense passive voice verbs more so than SNAs.

The excerpt below is from a space SNA grouped into Cluster 1. This SNA showed an especially high rate of present tense verbs (**bolded**) and modal verbs (underlined), two characteristic features of Cluster 1:

> **Scientists may be on the verge of finally unraveling mystery of antimatter**
>
> “We **are** kind of **tracing** the history [of hydrogen] in a historically parallel way - on the antihydrogen spectrum - to see if all these things **are** the same,” Hangst **says**. “It**’s** not the first measurement on antihydrogen, but it**’s** the first one of this character.” The more scientists **learn** about hydrogen, the more they will need to keep learning about antihydrogen. It **is** like chasing a moving target, Hangst **says**. Ten years ago, observing antihydrogen was not even possible. “This **is** really fundamental research at the limit of our understanding,” Hangst **says**. “We **are looking at** the most basic questions that you can ask about what **are** the symmetries of the universe. How **is** it actually? What **is** the structure of space and time?”
>
> Space SNA #219

As the excerpt shows, many of the present tense verbs were found in directly reported speech. In the four direct quotes, nine present tense verbs are used, two of which are progressive aspect. The fact that these verbs were present rather than past tense reflects the tendency for quoted material to discuss topics in the present time. Additionally, several of the present tense verbs are the copular *be*, which the interviewed researcher uses to define and characterize concepts or ideas.

The presence of quoted speech was only one reason for the high frequency of present tense verbs in Cluster 1. The following excerpt from an environment SNA has only one quotation but many present tense (**bolded**) and non-finite (underlined) verb phrases:

> **Let’s Face It: Going Green Is Simply No Match For Cutting Back on All The Junk We Buy**
>
> Being materialistic **does**n’t necessarily **make** a person feel happier, though. Or, for that matter, more miserable. But people who **buy** lots of things **do tend** to be less proactive with their finances. So even if wanting all the nice things actually **does**n’t **affect** our wellbeing, buying those expensive new shoes, latest PS4 games, and $20 bowls of acai and chia can reduce well-being if it **hits** the back pocket. By the same token, those who **reuse** what they can and **repair** what they can’t **tend** to feel good about saving their hard-earned dollars. “For very obvious reasons, if you **have** a proactive financial strategy and **put** money to the side and **live** within your means, it **has** positive well-being effects,” **says** Helm. Not that society **makes** this easy.
>
> Environment SNA #332

This excerpt, which comes from the second half of its respective article, shows a dense use of present tense lexical verbs and non-finite verbs, especially *-ing* participles. In the context of the larger SNA, these clauses are useful for elaborating on a logical argument, as most sentences include conjunctions, linking adverbials, or conditional adverbial clauses. The argument is elaborated with detail embedded in dependent clauses and its logical development made explicit with linking features, which together explain how consumer habits affect human psychology.

About one-third of texts in Cluster 1 were actually RAs. While there are likely several explanations, one of the most important appears to be the characteristics of the methods section. Simply put, if an RA’s methods section was absent, brief, or written in the active voice with some present tense verbs, then it was more likely to be grouped into Cluster 1. Consider the following excerpt from the methods of an environment RA (present tense, active voice verbs **bolded**):

Air pollution lowers Chinese urbanites' expressed happiness on social media

To create this cross-boundary spillover measure of local pollution, we **use** data from the Center for Earth System Science at Tsinghua University, which **has implemented** a series of improved, technology-based methodologies to develop an emissions inventory for China. Emissions are estimated by month and are captured at a spatial resolution of 0.25° × 0.25°. The inventory only **includes** anthropogenic emissions of PM2.5 and **does** not **include** secondary emissions from other pollutants, but this **is** acceptable for our research. We **construct** the instrumental variable NEIGHBOUR using the data for the year 201360. Here we **assume** the spatial variations of pollutant emissions and their monthly variations **are** stable between 2013 and 2014.

Environment RA #395

This excerpt, which originates from a methods section, adopts more present tense, active voice verbs than the stereotypical impersonal methods section. As a result, several sentences have the first-person plural pronoun *we* as their subject.

By contrast, RAs with prominent methods sections, especially those with many passive voice, past tense verbs were grouped into Cluster 2, which was defined by such features. While most Cluster 2 texts were indeed RAs with such methods sections, about one-third were SNAs. SNAs grouped into Cluster 2 read like brief summaries of their respective RA, with fewer present tense or modal verbs used to digress into the arguments, implications, and applications of research. Consider the following excerpt from the first half of a psychology SNA (past tense verbs **bolded**):

> **Study finds upper-class people attribute achievements to hard work when faced with evidence of class privilege**
>
> A series of experiments **recruited** hundreds of adult U.S. citizens from an elite West Coast university whose income **classified** them as upper middle-or upper-class. One group of participants **read** statements on either general inequity or class privilege's connection to greater educational opportunities. Afterwards, participants **answered** questions that **measured** their personal hardships in life. In addition, a separate group **read** about unearned advantages from people who make high incomes. Researchers **found** people who read about class privilege **reported** more life hardships than people who read about general inequity.
>
> Psychology SNA #241

Notably, no passive voice verbs can be seen in this excerpt, though there are plenty of past tense verbs. These verbs detail the steps taken by the researchers to carry out their study. In cases like this, the SNA conveys an unobtrusive style of objective summary. The SNA focuses almost entirely on the details on the study being reviewed, without little space for describing background literature, future directions, writer stance, or other digressions.

In short, this cluster analysis grouped the 800 RAs and SNAs into two clusters, each of which was defined largely by one register and its prototypical verbal features, such as passive voice past tense verbs in RAs and all other verbs in SNAs. Some texts from each register adopted non-prototypical verbal features, resulting in their being grouped in a different cluster. SNAs grouped into Cluster 2 read like impersonal summaries of the reported research, often using past tense verbs, while RAs grouped into Cluster 1 adopted more active voice and present tense verbs in their methods sections.

6.6 In focus: Modal verbs across topical categories

In section 6.4, modal verbs were shown to be significantly more frequent in SNAs. However, these features not only showed differences between register but also between topical domain. That is, a linear mixed effects model of best fit reported significant interactions between register and domain with a moderate effect size (R2c = .31) (see Appendix F.2 for the full output). Indeed, the patterns of modal verbs across register and domain were so divergent that interactional effects were found across all categories. The interaction plot in Figure 6.11 illustrates this point.

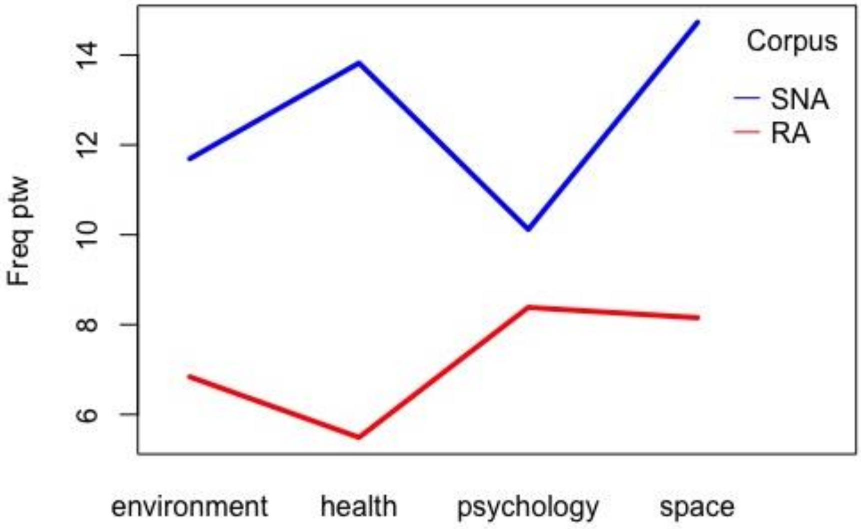


Figure 6.11 An interaction plot illustrating the mean normalized frequencies of modal verbs across register and topical category

This interaction plot illustrates the different patterns of modal verb use across register and topical domain. For example, psychology RAs used many modals relative to other domains, while psychology SNAs used the fewest among the other domains. A similar pattern is shown for health texts, where health RAs used the fewest modals among the domains but health SNAs used many modals. These divergences indicate an independence between register and domain with regard to the use of modal verbs.

When isolating topical domain, another trend can be seen. While all texts, regardless of register, relied most on possibility modals, the two nature-centric domains, namely space and environment, patterned more strongly with prediction modals relative to the human-centric health and psychology texts (Figure 6.12).

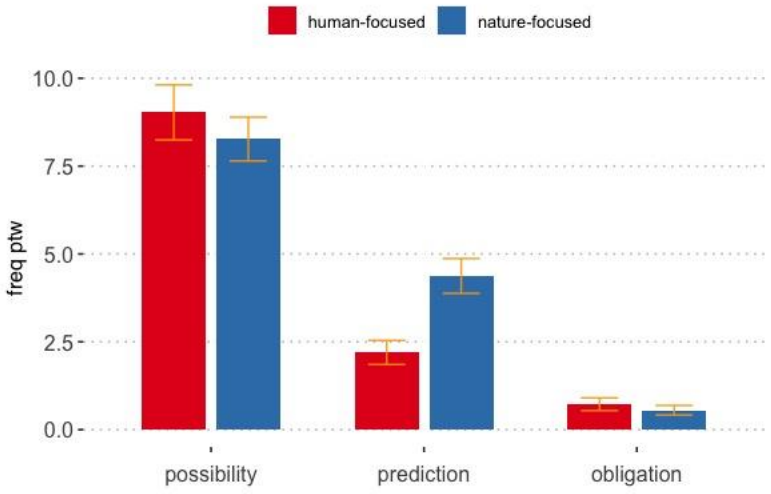


Figure 6.12 A bar graph illustrating rates of three modal verb categories by human-focused (health, psychology) and nature-focused (space, environment) SNAs

Figure 6.12 shows that the nature-centric texts, in either the RA or SNA corpora, used significantly more prediction modals than human-centric texts. Environment SNAs slightly preferred the prediction modal *will*, where it was often used to make confident predictions about the future of Earth's climate, particularly in relation to the deleterious effects of climate change (6.22) (*will* verbs **bolded**):

(6.22) Globally, we need to understand that changes of this nature **will** occur by 2100 and we need to plan how we are going to respond. [Environment SNA #305]

While environment texts preferred *will*, space SNAs slightly preferred *would* (see excerpt (6.20) for an example). Sentences with *would* often entertain hypothetical scenarios that engage the reader and match the often theoretical nature of many space texts. Compared to environment texts, it may be more difficult for space writers to make confident predictions about phenomena that occur in extraterrestrial settings, resulting in a preference for *would* over *will*.

6.7 Conclusion

In this chapter, I examined variation in the short verb phrase in the RA and SNA corpora. Specifically, I looked at variation in finite versus non-finite verb phrases, tense and aspect, grammatical voice, and modal verbs. I also presented a cluster analysis that groups texts together based on their use of these verbal features. SNAs were shown to use significantly more verbs overall, reflecting the tendency of academic writers to produce a nominal, informationally dense style of discourse, especially in the natural sciences (Biber & Gray, 2016). SNAs relied more heavily on present tense verbs, whereas RAs showed a comparatively greater reliance on past tense verbs. This difference can be partially explained by the greater presence of literature review and methodological description in RAs, where the past tense is often found. As the cluster analysis demonstrated, RAs with many past tense verbs also showed a greater use of passive voice. As Ding (2002) argues, the use of passive voice in research writing communicates to other scientists that a study is theoretically replicable to test and verify its findings. In SNAs, this function of passive voice is largely absent, and elaborated methodological description was described by informants as non-essential. Finally, modal verbs were found to be more frequent in SNAs, particularly the possibility modals *can*, *could*, and *might*, as well as the prediction modals *will* and *would*. Informants described using these features to hedge claims, which both protected writers from sensationalizing research results and communicated to readers that individual studies are often inconclusive on their own. Modal verbs were also unevenly distributed across topical domains and registers, showing different patterns of use across the domains within each register, suggesting that disciplinary influences do not necessarily translate from RA writing to SNA writing with regard to this feature.

7 REPORTING CLAUSES

In this chapter, I describe findings from an analysis of reporting clauses (RCs) across the RA corpus and SNA corpus. In particular, I seek to answer how the use of RCs differ across the registers in terms of overall frequency, the means by which they are introduced in texts, and the reporting verbs chosen. RCs were defined as clauses or sentences which attributed words, ideas, or other research-relevant information to sources other than the writer. Sentences or clauses with three patterns frequently used to attribute words and ideas were considered, namely TV+*that*Cl patterns, in which a *that* complement clause follows a transitive verb, *it*+*be*+TV+*that*Cl patterns, in which an extraposed *that* complement clause follows a passive voice transitive verb, and sentences with the complex preposition *according to*, which principally functions to identify a source of information (Biber et al., 1999, p. 871).

As an example, the following sentence is a TV+*that*Cl RC in which the subject (*the researchers*) is attributed the content of the *that* clause via the verb of communication *say*:

> The researchers, however, say the microbes could boost athletic performance. [Health SNA #125]

Following Charles (2006), I refer to the person or entity being attributed knowledge as the ‘author’, while the person who wrote the sentences is the ‘writer’. Thus, the author in the above example is *the researchers*, while the writer is the journalist who wrote SNA #125.

In section 7.1, I report on the frequencies of the three RC patterns across the registers. In 7.2, I describe four means by which writers instantiate RCs and report their use across the registers, as well as their associations with different grammatical contexts. Finally, I examine the authors of RCs in 7.3 before turning to their reporting verbs in 7.4.

7.1 Frequencies of RC patterns

In this section, I report on the frequencies of the three RC patterns. Figure 7.1 displays the normalized frequencies ptw of the three RC patterns across the registers. Because some patterns were much more frequent than others, the exact normalized frequencies are displayed on the graph for enhanced clarity.

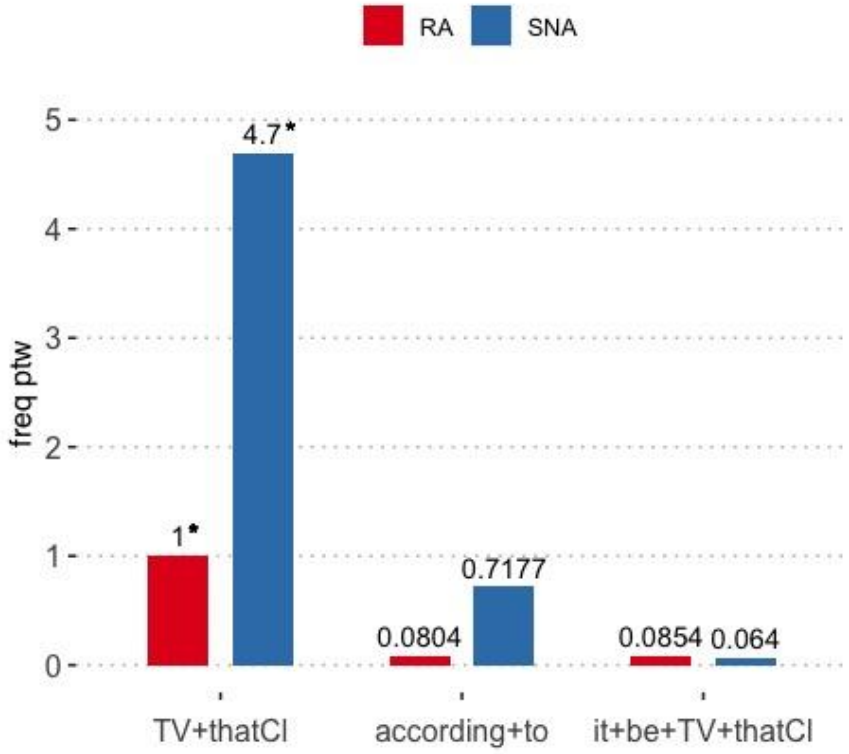


*Figure 7.1 A bar graph illustrating the normalized rates of three RC patterns across the registers. *Based on partial analysis of all TV+*that*Cl patterns*

Figure 7.1 illustrates that SNA writers employed RCs more frequently than did RA writers. Overall, SNA writers used the three patterns 5.48 times ptw compared to RA writers' 1.17 times ptw. Between the three patterns, TV+*that*Cl RCs were most frequently used for attribution, which is intuitive considering that non-extraposed *that* clauses are much more frequent than extraposed ones (Biber et al., 1999, p. 670). TV+*that*Cl RCs were particularly frequent in the SNA corpus, where they were used more than four times as frequent than in the RA corpus.

By contrast, RAs showed a slightly higher rate of *it+be*+TV+*that*Cl RCs compared to the SNA corpus, which is likely related to two points. First, as Chapter 5 and 6 demonstrated, passive voice verbs were generally more frequent in RAs, resulting also in the higher rate of *it+be*+TV+*that*Cl RCs. Second, patterns involving extraposed *it* clauses tend to convey an impersonal and objective voice (Hewings & Hewings, 2002; Zhang, 2015), a style of writing preferred in academic writing.

Finally, *according+to* RCs were more frequent in the SNA corpus relative to the RA corpus. While this finding may reflect a greater use of prepositional phrases functioning as adverbials in SNAs (cf. section 5.5), it may also suggest a greater density of linguistic features used to attribute knowledge in SNAs. As one informant described,

> *In journalism, I'm trying to make it clear that I'm not just making this stuff up, and that I am getting it from a source whether that is the research article or the researchers or whatever. That is coming from there. [Informant #6]*

In other words, SNAs attribute knowledge at a higher rate in general, in part because writers want to add credibility to their articles, whereas a significant portion of RAs is dedicated to the unique contribution that the authors themselves add to an academic debate. The higher rate of *according+to* in the SNAs may also suggest that this pattern is more common in news writing than academic writing.

7.2 Frequencies of RC methods and their associations with grammatical contexts

In RAs, a common marker of attribution is in-text citation, which signifies that the content of the sentence is to be attributed to a source other than the writer. Studies of attribution in RAs often rely on in-text citations to identify RCs (e.g., Hyland, 2002). However, SNAs rarely adopt academic-style citation, so additional methods for identifying RCs were developed. These methods are described below.

Deixis. This method includes the use of nouns or pronouns to refer back to previous RCs without the use of citation. These instances occur anywhere but the start of a chain of reference. Hence, the authors of these sentences only make sense in the context of a previous instance of attribution. Deixis often includes common nouns determined by definite determiners (e.g., *<u>The study</u> shows the potential for mapping individual trees worldwide…*), third person pronouns (e.g., *In short, <u>they</u> argue that there is no such thing as a "steady state" for an ecological…*), or other common nouns referring to previously established research, regardless of determiner (e.g., *<u>Results</u> showed that accelerated epigenetic age was related to lower IQ scores…*).

Citation. This method includes the use of integral or non-integral citations to attribute a proposition to someone or something. In addition to integral and non-integral citation, a third sub-type called 'initial citation' was introduced for this analysis. Like integral citations, initial citations have a syntactic role in the clause, but the name of the author is replaced by a common noun, usually referring to a single study (e.g., *<u>A new study</u> published in Nature Astronomy further investigates Enceladus' possible habitability…*). They generally occur first in a chain of reference, which is why they are called initial citation, and later are cited with fuller detail.

General reference. This method refers to RCs that attribute propositions to a non-specific group of people or entity, such as scientists, studies, or schools of thought, and are often plural common nouns (*However, <u>scientists</u> think they too could soon be gone…*), extraposed *it* clauses (*It has been acknowledged that childhood traits and circumstances can have…*), or the inclusive *we* (*<u>We</u> know that one strain of bacteria in our gut can effectively convert…*). General reference differs from other categories in that it is not possible to identify a specific author and a citation is not present.

Quotation. Finally, this method reports the exact words of an author using quotation marks. Based on a qualitative review, I determined that a minimum of 5 words was required to be considered a quotation, a guideline which helped avoid including short, quoted phrases that were not intended to be attribution. Sentences comprised entirely of a quotation (*"We know that inflammatory pathways induced by infection force blood stem cells…"*), as well as sentences with only partial quotations (e.g., *Women in the study "had benefit beyond relief of pain"*) were included.

When categorizing RCs, an order of operations was adopted. Quotation superseded all other categories, followed by citation and then the deixis and general. For example, if an RC included a quotation, citation, and general reference all in one sentence, it would be coded as quotation. However, if no quotation was present, it would then be coded as citation. If neither quotation nor citation were present, leaving only a general reference, it would be coded as general reference. General reference and deixis are mutually exclusive and thus did not conflict with one another.

7.2.1 Frequencies of RC methods

Figure 7.2 displays the proportional use of the four RC methods, namely deixis, citation, general reference, and quotation, across the registers.

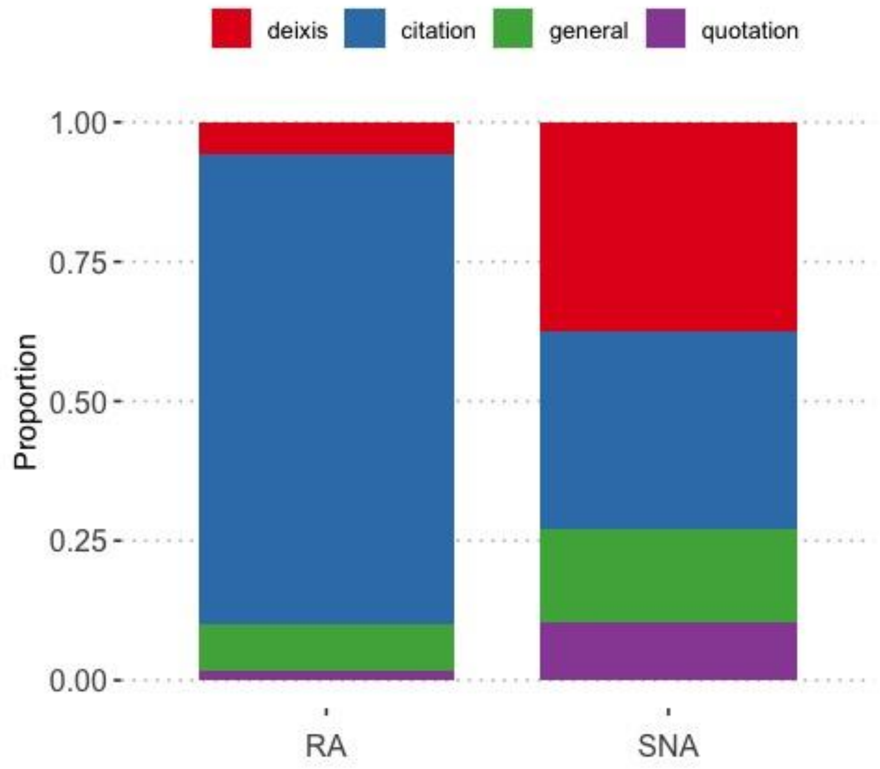


Figure 7.2 A bar graph illustrating the proportional use of four RC methods across the registers

Figure 7.2 displays divergent profiles of RC methods between the registers. In the SNA corpus, both deixis and citation were about equally preferred for attribution, while general reference and quotation were relied on to a smaller, though not insignificant, extent. In short, all four methods played a part in attribution. In the RA corpus, deixis, general reference, and quotation were used very sparingly, with most RCs employed academic-style citation. These preferences relate to the situational characteristics of both registers. RAs rely on academic-style citation because researchers are expected to situate their study within a body of research, which should be cited in-text to increase credibility and provide expert readers the information needed to seek out the cited work if desired. SNAs, by contrast, are short texts that highlight only brief moments in an academic conversation. Thus, they (largely) introduce a single RA to be reported on and reference largely that RA, occasionally providing broader context through less precise general references.

The reliance on citation in RAs reflects the most mentioned method of attribution in academic writing, academic-style citation. Excerpt (7.1) below exemplifies an instance of integral citation, where the author's name is written in the prose of the sentence (citation **bolded**).

(7.1) According to **Fatima et al. [66]**, palmitic acid is an intracellular signaling molecule involved significantly in disease development. [Health RA #148]

While it may appear odd that SNAs also showed a moderate reliance on citation, this reliance was due to the 'initial citation' category described in the start of this section, rather than academic-style citation like (7.1). Consider the following excerpt (citation **bolded**):

(7.2) However, **a recent study by Newcastle University** suggests that this four percent makes up almost a fifth of microplastics in our oceans. [Environment SNA #374]

This initial citation introduces the study that will be reviewed throughout the SNA. These RCs often come near the beginning of an SNA, using language like "a recent study," "new research," and similar phrases. Thus, a key difference between initial citation and academic-style citation is that the latter is often used to contextualize new research within a larger body of literature, while the former introduces a single study without developing the same kind of context.

By this same logic, deixis was common in SNAs but not RAs. Once a new study was introduced via initial citation, the SNA writer could then continually refer back to it with deictic RCs, such as (7.3) (deixis RC **bolded**).

(7.3) **The study** found that emotional intelligence was linked to better times in a half-marathon race. [Psychology SNA #215]

The subject of this RC is the definite noun phrase *the study*, which is definite because it refers to an already-introduced RC.

That deictic RCs are relatively frequent in SNAs relative to RAs suggests differences with regard to the ways that the registers perform attribution. In RAs, RCs are often cumulative. Writers attribute words and ideas to numerous different authors, as the ever-growing body of literature across academia demands researchers to situate their study within a larger and larger conversation (Hyland & Jiang, 2019). However, RCs in SNAs are repetitive. Writers attribute words and ideas to the same authors iteratively, contributing knew information but attributing it to the same authors. One informant described this difference this way,

> [In science news] *we're not trying to argue to a community and convince a community "Why is this advancing knowledge?" We're just telling that it has. This is now new, it's interesting, and it is something we didn't know before. The history part of it—unless it's really going to help, we don't need to actually tell you too much about that.* [Informant #2]

This informant argues that contextualizing new research in a longer academic discussion in a detailed manner is not essential for SNAs and their readers. Rather, writers craft SNAs to be brief and emphasize the novelty of research findings. Arguably, this relates to the lower variety of authors in the RCs of SNAs.

The remaining two methods, general reference and quotation, were relied upon more by SNA writers, especially quotation. RA writers rarely quote others because it is not economical and can only be attributed to one author at a time. While for SNA writers, quotes can add a perspective on research that the published RA alone cannot. Some quotes resemble the "contingent repertoire" described by Gilbert and Mulkay (1984). The researchers interviewed biologists and compared their informal accounts of research activities with their published accounts in RAs. While RA writing presents activities in an impersonal and empirical style, the scientists' informal accounts depict their actions as motivated by personal motivations and inclinations. To see this discourse style in an SNA, consider the following excerpt (7.4):

> *(7.4)* "I think that is the ultimate question people want answered, but I became more and more interested in how each mass shooting differed and what factors might have contributed to these variations. I asked myself the simple question, 'are the factors associated with variations in violence across shootings similar to those which predict them?' I also wanted to know if certain policies might be helpful in addressing these factors. Thus, this research began." [Psychology SNA #239]

This excerpt is one long direct quotation given in an interview with the SNA's writer. In comparison with the typical style of research writing (what Gilbert and Mulkay call the 'empiricist repertoire'), the quoted researcher here frequently uses first person pronouns, adopts mental verbs positioning the content as the result of their subjective thought process, and describes the motivations behind the research. In other cases, quotations were brief and did not engage the same repertoire but rather a more typical empiricist repertoire.

SNA writers also used more general references, which were principally carried out with TV+*that*Cl patterns in the SNA corpus but *it*+*be*+TV+*that*Cl patterns in the RA corpus. Informants described its usefulness in a variety of ways. For example, general reference was said to convey a sense that the attributed content is factual and holds true for an indefinite amount of time (Informant #4). Additionally, these RCs were described as being more economical

(Informant #2, Informant #5), useful for the accomplishing the expected brevity of SNAs. Finally, general reference was also said to reflect the casualness of SNA writing relative to academic writing:

> *When you build a theoretical background in a scientific article, you have to be very precise when you make that statement [...] you have to support it with the reference about where you read it and who stated that. In news articles, there is no such requirement [...] you can pretty casually put things like "researchers", "everyone", "previous studies", "previous research" blah blah blah.* [Informant #1]

The association between RC sub-types and grammatical context is explored next, in section 7.2.2.

7.2.2 The association of RC methods with different grammatical contexts

The four RC methods were also examined for their associations with particular grammatical contexts, namely whether they preferred active or passive voice, past or present tense, or simple or non-simple aspect. I begin with grammatical voice before turning to tense and finally aspect. The raw frequencies used to compute all statistical tests can be found in Appendix G.

To test whether any significant associations exist between the four methods and grammatical voice, separate 2×4 Chi-square tests of independence were computed for each register. The χ^2 statistic was significant for both the RA corpus (χ^2=43.12, p < .001, Cramer's *V* = 0.20) and SNA corpus (χ^2=43.12, p < .001, Cramer's *V* = 0.15), reporting medium effect sizes for both. To visualize these associations, mosaic plots for each test are displayed in Figure 7.3.

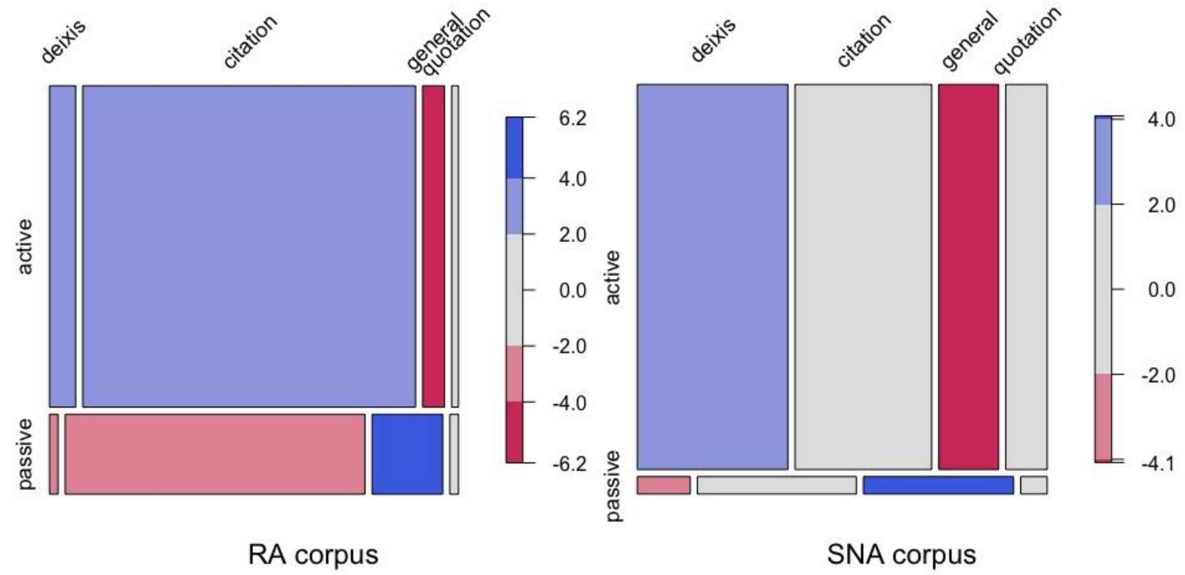


Figure 7.3 Mosaic plots of four RC methods in active and passive voice across two registers

Along the left-hand side of the mosaic plots are active and passive voice, while the four methods are along the top. The colors correspond to the strength of the cells' respective adjusted standardized residuals, where dark red colors correspond with large negative residuals and dark blue colors with large positive residuals. Colored areas indicate significant associations, while gray areas suggest no significant associations. The overall sizes of the boxes represent the size of the raw frequencies of their respective cells.

Figure 7.3 reveals similarities and dissimilarities between the registers. Deixis was positively associated with active voice and negatively associated with passive voice in both registers. Similarly, general reference was strongly positively associated with passive voice and strongly negatively associated with active voice in both registers. One dissimilarity, however, was the positive association between citation and active voice in the RA corpus, which was not reciprocated in the SNA corpus.

Next, I tested the associations between RC method and past or present tense. These tests resulted in significant statistics for both the RA corpus (χ^2=22.73, p < .001, Cramer's *V* = 0.14) and SNA corpus (χ^2=29.71, p < .001, Cramer's *V* = 0.18), with medium effect sizes for each. To visualize these associations, mosaic plots are displayed in Figure 7.4.

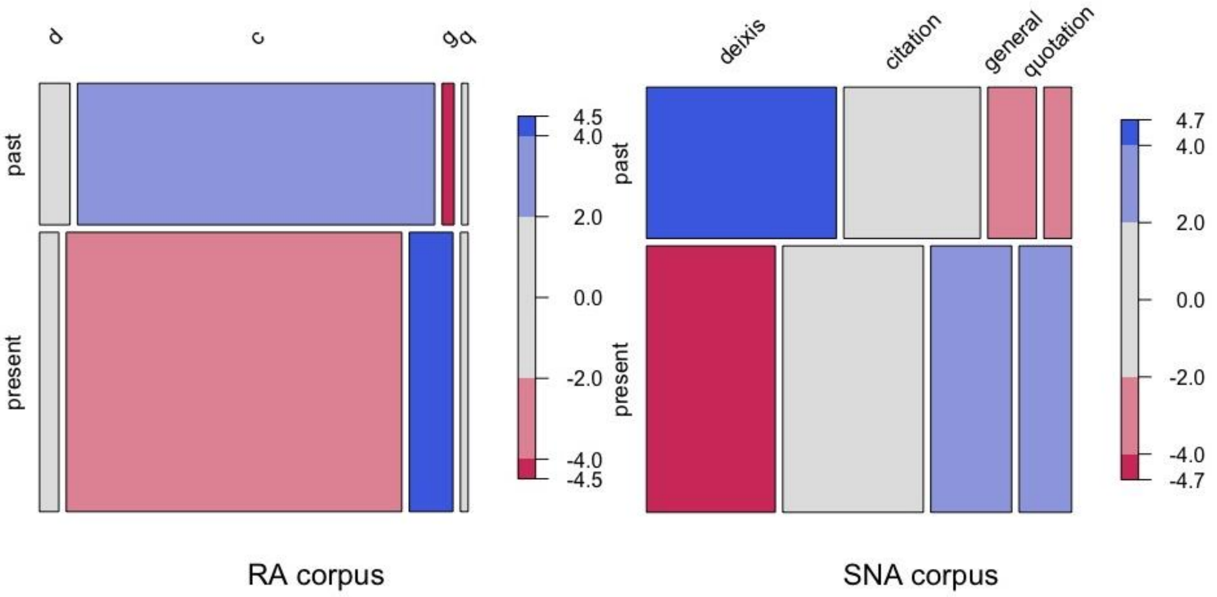


Figure 7.4 Mosaic plots of four RC methods in past and present tense across two registers. Note that labels for the RA corpus's mosaic plot have been truncated for readability.

Figure 7.4 shows some dissimilarity between the registers. First, deictic RCs were strongly positively associated with the past tense and negatively associated with the present tense in the SNA corpus but showed no such association in the RA corpus. Additionally, quotation was negatively associated with the past tense and positively associated with the present tense in the SNA corpus but not the RA corpus. Finally, only the RA corpus showed a positive association between citation and past tense and negative associated between citation and present tense. One similarity, however, was in the positive association between present tense and general reference, as well as its negative association with past tense, which was found in both registers.

Finally, I tested the associations between RC method and simple or non-simple aspect, resulting in significant statistics for both the RA corpus (χ^2=20, p < .001, Cramer's *V* = 0.13) and SNA corpus (χ^2=75.41, p < .001, Cramer's *V* = 0.27). The effect size for the RA corpus was near-medium and the effect size for the SNA corpus was near-large. To visualize these associations, mosaic plots are displayed in Figure 7.5.

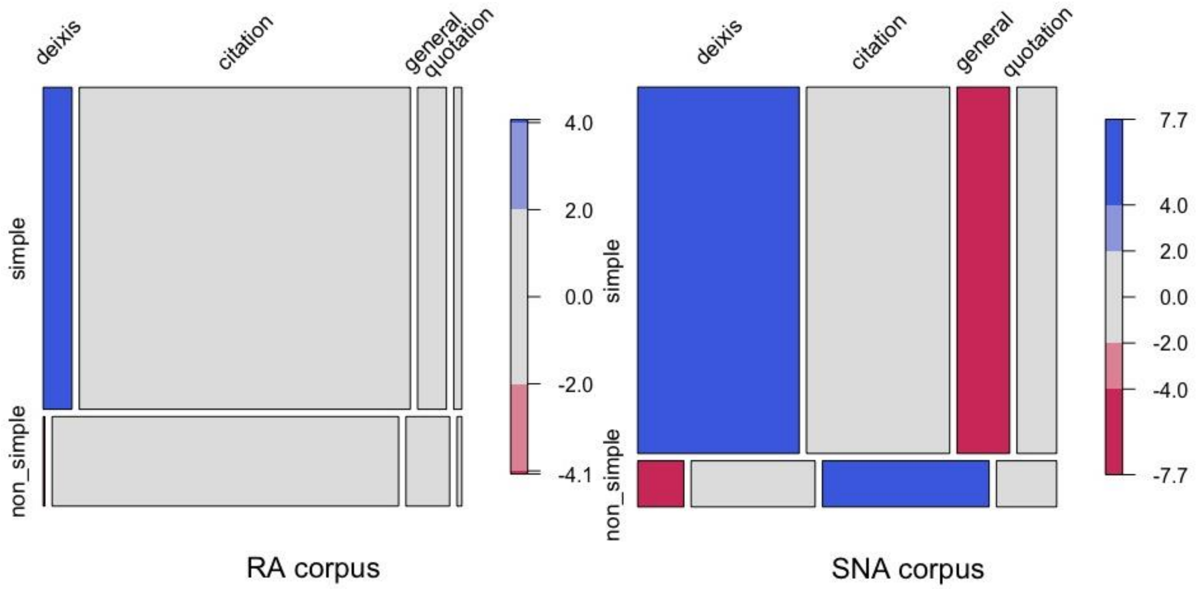


Figure 7.5 Mosaic plots of four RC methods in simple and non-simple aspect across two registers

Figure 7.5 shows that deixis was positively associated with simple aspect and negatively associated with non-simple aspect in both registers. On the other hand, general reference strongly positively associated with non-simple aspect and negatively associated with simple aspect in the SNA corpus only.

The noted statistical associations suggest trends in the grammar of the four RC methods. First, both deixis and general reference showed several preferences for certain grammatical contexts. Deixis associated with simple past tense verbs in the active voice. These sentences function to identify the primary findings of a study introduced earlier in the text. Consider the following excerpts (simple past tense verbs **bolded**):

(7.6) The findings, published in the Journal of Psychosomatic Research on March 17, **showed** that Pennsylvania mothers who experienced PTSD after giving birth were twice as likely to score in the bottom third of child bonding evaluations. [Psychology SNA #161]

(7.7) They also **found** that these recently exposed, subglacial bryophytes can regenerate, which may have important implications for recolonization of polar landscapes. [Environment RA #306]

In (7.6) and (7.7), the subject is a definite noun phrase or pronoun linking back to a previously introduced study, the main verb is in the simple past tense, and a complement clause unpacking a study's findings is attributed to the subject. Thus, despite register differences, this type of attribution often functioned similarly.

General reference was associated with present tense verbs in the passive voice. In the SNA corpus, some of these sentences also adopted the perfect aspect. Consider the following excerpts (present passive verbs **bolded**):

(7.8) It **has** even **been suggested** that A. muciniphila supplements could enhance the efficacy of certain cancer treatments. [Health SNA #138]

(7.9) Finally, it **is recognized** that BMI is a relatively poor indicator of the percent of body fat. [Psychology RA #209]

As these excerpts illustrate, the *it*+*be*+TV+*that*Cl pattern was partially responsible for the strong association between general reference and passive voice, since this pattern is obligatorily passive. However, Figure 7.2 also showed that general reference was more frequent in SNAs, due in part its use in TV+*that*Cl patterns there (7.10) (general reference **bolded**).

> *(7.10)* So **some experts** caution that while creating more active, appealing playgrounds may be a step in the right direction, getting kids to move more requires a multipronged approach. [Health SNA #4]

Here, the subject is a vague group of *experts* who, according to the writer, cautions against a simple solution to a challenging problem. Unlike (7.8), excerpt (7.10) is in the active voice, utilizing the TV+*that*Cl pattern, a grammar pattern which the RA corpus rarely used to attribute knowledge with general reference.

The remaining two sub-types, citation and quotation, showed fewer associations with particular grammatical contexts. In the RA corpus, citation associated with the past tense and active voice, functioning similarly to excerpts like (7.7). Such RCs were considerably rarer in SNAs. For one example, consider the following excerpt from a psychology SNA (past tense verbs **bolded**):

> *(7.11)* A previous study **showed** that the most common misconception about bisexual people is that they are in denial about their true orientation. [Psychology SNA #188]

(7.11) is similar to past tense citations in RAs in that the RC explains the findings of an individual study that has not been previously introduced in the text. However, unlike those in RAs, it lacks the information typical of academic-style citation, such as the names of the researchers and year of publication.

7.3 The authors of RCs

In this section, I examine the authors of RCs in more detail. Because the authors of *it*+*be*+TV+*that*Cl RCs are not explicitly stated, this pattern was not included in the analysis. The authors of TV+*that*Cl RCs are the grammatical subjects, while the authors of *according+to* RCs are the noun phrase complements of the preposition *according to*. Both kinds of authors are included in the analyses below.

First, I explore the relationship between register and author animateness, specifically whether an author refers to a human or not (Table 7.1).

*Table 7.1 Frequencies for how many authors of TV+*that*Cl and* according+to *RCs referred to humans or non-human entities*

	RA corpus		SNA corpus	
	Raw	Proportion (%)	Raw	Proportion (%)
Human author	323	34%	607	61%
Non-human author	638	66%	394	39%
Total	961	100%	1001	100%

Table 7.1 reveals different trends for the registers. SNAs use human authors in about 2/3rds of RCs, while RAs use non-human authors in about 2/3rds of its RCs. The association between the humanness of the author and register is a significant one (χ^2=142.58, p < .001) with a medium effect size (Cramer's *V* = 0.27). In both registers, the *according+to* pattern attracted the highest proportion of non-human authors, while TV+*that*Cl patterns were most amenable to having either human or non-human authors.

Table 7.2 presents the most frequent authors of TV+*that*Cl and *according+to* RCs.

*Table 7.2 The ten most frequent authors of TV+*that*Cl and* according+to *RCs across the registers*

RA corpus		SNA corpus	
author	raw freq	author	raw freq
study	174	study	133
[name] *et al.*	132	**researcher**	125
research	57	research	66
model	41	**team**	64
work	31	author	43
they	28	they	42
data	23	**scientist**	40
analysis	20	**we**	40
evidence	19	**result**	27
author	16	analysis	16

Note: **bolded** words are those that are unique to their respective column

Table 7.2 shows that the primary human author in the RA corpus is any group of authors followed by *et al.* (e.g., *Liu et al.*). *They* and *author* were also attested among the top ten authors in the RA corpus. *They* reflects the fact that contemporary academic research is largely conducted by multi-author teams (Hyland & Jiang, 2019). It also shows that RA writers tended to insert several RCs in a small space without belaboring one study at length (7.12) (RC authors **bolded**).

(7.12) **Greaves & Rice (2010)**, **Williams (2012)**, and **Najita & Kenyon (2014)** showed that only a tiny fraction of disks contain enough mass to explain the mass of the large

number of observed gas giant planets in the standard core accretion model. **They** proposed that planet formation must have been underway by the time these disks were observed (1/21-3 Myr). Similarly, **Mulders et al. (2015, 2018)** used […] [Space RA #134]

(7.12) shows three instances in which the RA writer attributes propositions to multiple studies in the same RC. By doing so, RA writers can generalize knowledge across multiple citations, saving space and allowing one to comment on the literature more easily. While such a function is theoretically possible for SNA writers, they appear to prefer the use of general reference to comment on a wide swath of research (7.13) (RC author **bolded**).

(7.13) But **astronomers** think many of them—in fact, the vast majority from the early universe—may be in hiding, camouflaged behind much closer galaxies. [Space SNA #62]

As informants noted in section 7.2.1, general references like (7.13) are more economical and more casual than the elaborate literature reviews in many RAs. Moreover, writers are often expected to produce SNAs in a short period of time, so general reference would seem a more plausible method for summarizing across previous literature or knowledge than the otherwise complex task of narrating a body of literature with detail.

Other authors in the RA corpus were largely inanimate, referring to various research objects like *model, work, data,* and *evidence*. Because RAs utilize the citation style of attribution most often, even RCs with non-human authors typically had a non-integral citation pointing readers toward a source of further information (7.14) (RC author **bolded**, non-integral citation underlined).

(7.14) Emotional disclosures seem to foster a more intimate relationship with the follower, according to **Reis & Shaver's interpersonal process model of intimacy** (Hassan et al., 2016; Reis & Shaver, 1988). [Psychology RA #175]

In the SNA corpus, a greater variety of human authors can be seen, such as *researcher, team, author, they, scientist,* and *we*. These authors were often used deictically. An initial citation is used to initiate a chain of reference, after which human authors determined by *the* can be used to refer back to the initial citation (7.15) (RC authors **bolded**).

(7.15) […] **a recent study** published in the Proceedings of the National Academy of Sciences suggests that at least one of Earth's past mass extinctions might have been the result of a nearby supernova. […] **the scientists** behind this latest study found plant spores that were burned by ultraviolet light, which suggests they were around when the ozone was depleted. [Space SNA #38]

In (7.15), the SNA writer introduces the study to be reported on with an initial citation (*a recent study…*) and later refers back to it using the noun phrase *the scientists behind this latest study*.

By contrast, other than *study,* there were fewer authors referring to research objects in the SNA corpus. In the RA corpus, *study* was often used in its plural form, *studies*, used to generalize across multiple citations. In the SNA corpus, however, it was almost entirely used in the singular form and determined by *the*, serving as a cohesive tie in a chain of reference.

7.4 The reporting verbs of RCs

In this section, I examine the main verbs of RCs, also known as reporting verbs (Thompson & Ye, 1991). Analysis is restricted to two patterns, namely TV+*that*Cl and *it+be*+TV+*that*Cl patterns, because the author in *according+to* patterns is not the grammatical or semantic subject of the main (i.e., reporting) verb. In 7.4.1, I examine the associations of three semantic categories of reporting verbs and register, before examining each category in more detail.

7.4.1 The association of reporting verb semantic categories and register

Table 7.3 displays the raw frequencies and proportional use of three reporting verb semantic categories across the registers.

*Table 7.3 Frequencies of reporting verbs in TV+*that*Cl and* it+be+*TV+*that*Cl RCs by semantic category and register*

	RA corpus		SNA corpus	
	raw freq	proportion (%)	raw freq	proportion (%)
Discourse acts	413	43%	336	42%
Research acts	480	50%	316	39%
Cognition acts	73	7%	149	19%
Total	966	100%	801	100%

Table 7.3 shows that both registers relied most heavily upon discourse acts and research acts. While fewer in number in both registers, cognition acts were more frequent in the SNA corpus. In fact, the difference in use of cognition acts between the register has an odds ratio of 2.8, meaning that the odds of coming across a cognition act is 2.8 times higher in the SNA corpus. The RA corpus made greater use of research acts, resulting in an odds ratio of 1.52. Both registers made use of discourse acts to a similar extent. Below, these three categories are examined in more detail.

7.4.2 Discourse acts

Discourse acts frame attributions as spoken or written expressions made by the author. As Table 7.3 showed, these verbs were relied upon to a similar extent in both registers. Both corpora used about 30 different verbs to produce about 43% of all reporting acts in each register. Table 7.4 displays the most frequent discourse act verbs in both registers.

Table 7.4 The ten most frequent discourse act verbs across the registers

RA corpus		SNA corpus	
verb	raw freq (%)	verb	raw freq (%)
suggest	173 (42%)	suggest	107 (32%)
indicate	54 (13%)	**say**	77 (23%)
argue	43 (11%)	note	23 (7%)
report	38 (10%)	report	14 (4%)
propose	29 (7%)	indicate	13 (4%)
note	12 (3%)	argue	12 (3%)
state	9 (2%)	**explain**	10 (3%)
point out	7 (2%)	point out	9 (2%)
document	6 (1%)	**tell**	9 (2%)
recommend	5 (1%)	**add**	7 (2%)

Note: **bolded** words are those that are unique to their respective column

Table 7.4 shows that many of the most frequent discourse acts, such as *suggest, indicate,* and *argue*, were used in both registers. However, the verbs *propose, state, document,* and *recommend* were specific to the RA column, whereas *say, explain, tell,* and *add* were specific to the SNA column.

Since *suggest* was the most frequent reporting verb in either register, it is worth exploring this verb further. Its typical use was similar across the register, where it was used to make tentative suggestions about the interpretation of research results (7.17):

(7.17) He and his colleagues **suggest** that many people may be unaware that smoking marijuana affects the body differently from eating the drug. [Health SNA #8]

Although this reporting verb was frequent in both registers, SNA writers relied upon it to a lesser extent, likely because of its typical subjects. Whereas verbs like *say*, *note*, and *report* had largely human subjects, about 3/4ths of RCs with *suggest* had subjects like *research, results,* and *findings*. Thus, since SNAs largely used RCs with human authors (see section 7.3), SNA writers may have relied less upon *suggest*.

Following *suggest*, the second most frequent discourse act in SNAs was *say*. Tellingly, *say* did not occur often with quotation RCs. Instead, it was mostly used with the citation and deixis methods (Table 7.5).

Table 7.5 The occurrence of the discourse act say *across four RC sub-types in the SNA corpus*

Sub-type	Raw freq
Citation	41
Deixis	25
General	5
Quotation	8
Total	79

In other words, SNA writers often used *say* to indirectly report the words of researchers. Excerpt (7.18) illustrates a deictic use of this reporting verb (**bolded**):

(7.18) The team **says** they hope to begin human clinical trials of the new contraceptive drug delivery system within the next three to five years. [Health SNA #233]

Say has practical utility for SNA writers, who often interview scientists and therefore need to report their words in fairly transparent terms. However, *say* is motivated by other factors as well. One informant highlighted that they must demonstrate that their information comes from expert sources and is not "made up" (Informant #6). Thus, using *say* with indirect reports may serve to communicate that the content of the sentence is reliable, originating from an expert despite not being directly quoted. Another informant noted that alternating between direct quotation and paraphrase makes the writing less repetitive (Informant #4).

Notably, *say* also withholds writer stance, communicating no clear indication of neither the writer's interpretation nor the author's. This explanation reflects a larger trend in the use of discourse acts by SNA writers. Table 7.6 displays the most frequent discourse act verbs by the kind of evaluation that they imbue upon the writer or author as demonstrated in previous literature (see Charles, 2006; Hyland, 2002b, 2007; Thomas & Hawes, 1994; Thompson & Ye, 1991).

Table 7.6 The top ten most frequent discourse acts verbs in both registers grouped by the kind of evaluation indicated by the verb

	RA corpus		SNA corpus	
	verbs	proportion (%)	verb	proportion (%)
Tentative	Suggest, indicate, propose	62%	Suggest, indicate	36%
Neutral	State, report	12%	Say, report, tell	29%

Certain	Argue, note, point out, recommend, document	26%	Argue, explain, note, point out, add	35%

Table 7.6 shows that RA writers adopted tentative reporting verbs like *suggest*, *indicate*, and *propose* in about 62% of all discourse acts, while the same verbs comprised just 36% of discourse acts in the SNA corpus. By contrast, SNA writers relied more upon discourse acts expressing certainty and neutrality, especially *say*. For example, (7.19) below illustrates the use of the more certain reporting verb *explain*:

(7.19) Researchers **explain** that while these developments are of course culturally important, they can have a negative effect on the forest, and they recommend that such further developments be done with conservation in mind. [Environment SNA #359]

The use of *explain* here positions the writer as being confident that the explanation is truthful and reliable. In addition, *explain*, in particular, helps facilitate reader understanding, as it expounds upon a topic or idea by providing an explanation (Calsamiglia & Ferrero, 2003).

In short, the RA corpus is noteworthy for its reliance upon tentative discourse acts and dearth of neutral acts, while the SNA corpus is noteworthy for its more even reliance on tentative, neutral, and certain discourse acts.

7.4.3 Research acts

Research acts, relied upon most by RA writers, refer to the mental or physical processes that relate to research activities, usually study procedures or findings. Despite having the fewest number of unique verbs, research acts comprised the greatest proportion of overall reporting verbs. Table 7.7 displays the most frequent research acts in the registers.

Table 7.7 The ten most frequent research act verbs across the registers

RA corpus		SNA corpus	
verb	raw freq (%)	verb	raw freq (%)
show	187 (39%)	find	142 (45%)
find	149 (31%)	show	79 (25%)
demonstrate	27 (6%)	reveal	15 (5%)
estimate	26 (5%)	conclude	14 (4%)
conclude	24 (5%)	estimate	14 (4%)
reveal	19 (4%)	**discover**	11 (3%)
predict	13 (3%)	predict	10 (3%)
observe	10 (2%)	demonstrate	8 (3%)
hypothesize	6 (1%)	hypothesize	6 (2%)
confirm	3 (1%)	confirm	4 (1%)

Note: **bolded** words are those that are unique to their respective column

Table 7.7 shows that the verbs *find* and *show* were responsible for the majority of research acts in both corpora. Between the verbs, *show* was preferred by RA writers and *find* by SNA writers. While both verbs function to report research findings, *show* communicates a positive writer stance toward the proposition while *find* a neutral one (Hyland, 2002b). That is, whereas *show* indicates that the writer accepts the scientist’s findings as factual, *find* identifies a study’s conclusion without committing to its truth value.

Thus, SNA writers may rely more on *find* because it appeals to the journalistic ideal of objectivity. Science is contested terrain, so science communicators often default to neutrality when unsure of the truth value of certain claims (Dunwoody, 2014, p. 33). *Find* also fits into a small group of research acts that arguably convey a ‘discovery’ characteristic of academic research. That is, reporting verbs like *find*, *reveal*, and *discover* position study findings in a more dynamic and intriguing way, making them more attractive to SNA writers and possibly their readers. Consider the frequencies of these verbs across the corpora (Table 7.8).

Table 7.8 Frequencies of three 'discovery' reporting verbs across the registers

RA corpus		SNA corpus	
verb	raw freq. (%)	verb	aw freq. (%)
find	149 (31%)	find	142 (45%)
reveal	19 (4%)	reveal	15 (5%)
discover	2 (<1%)	discover	11 (3%)

Table 7.8 shows that these 'discovery' verbs comprise about 53% of research acts in the SNA corpus but about 35% in the RA corpus, suggesting research acts in SNAs more often present study findings as discoveries. Discovery verbs like *find* also prefer human authors, while other common research acts like *show* prefer non-human authors. Consider the (7.20) and (7.21) below (reporting verbs **bolded**).

(7.20) In fact, biologists have **found** that a previously unstudied population of polar bears (Ursus maritimus) in the Chukchi Sea, between Alaska and Russia, is actually thriving. [Environment SNA #323]

(7.21) The method used has been **shown** that may overestimate the column density through which the disc gas cools (Wilkins & Clarke 2012; Young et al. 2012; Lombardi, McInally & Faber 2015) by a factor of a few. [Space RA #256]

It is difficult to identify the precise ways in which *find* in (7.20) communicates a more neutral stance than *show* in (7.21). In fact, the use of *in fact* in (7.20) would seem to position the writer as fairly confident in the sentence's proposition. However, a more easily observable feature of these two verbs is the former's preference for human authors, leading to a preference for these verbs in SNAs, and the latter's preference for non-human authors, leading to a preference for these verbs in RAs. Moreover, use of discovery verbs with human authors more easily allows SNA writers to craft the kinds of texts they set out to—to tell stories of science involving researchers, their actions, their words, and their beliefs.

7.4.4 Cognition acts

Cognition acts, which were relied upon most by SNA writers, refer to mental activities (e.g., *think*, *believe*, *know*) not considered to be part of the research process (e.g., *find*, *hypothesize*, *calculate*). Table 7.9 displays the most frequent cognition acts in both registers.

Table 7.9 The ten most frequent cognition act verbs across the registers

RA corpus		SNA corpus	
verb	raw freq (%)	verb	raw freq (%)
know	18 (23%)	think	42 (28%)
assume	14 (19%)	believe	29 (20%)
believe	7 (10%)	know	21 (14%)
recognize	6 (8%)	**hope**	12 (8%)
accept	4 (5%)	**suspect**	9 (6%)
infer	4 (5%)	assume	6 (4%)
speculate	3 (4%)	**realize**	6 (4%)
think	3 (4%)	**agree**	3 (2%)
consider	2 (3%)	expect	3 (2%)
expect	2 (3%)	**learn**	3 (2%)

Note: **bolded** words are those that are unique to their respective column

Table 7.9 shows that the verbs *think* and *believe* were responsible for a large portion of cognition acts in the SNAs, whereas *know* and *assume* were the most frequent cognition acts in the RAs. While it has been argued that *think* positions the author as having a positive stance toward a proposition and *believe* a more tentative one (Hyland, 2002b), it was often difficult to find such a distinction in the corpora. Consider excerpts (7.22) and (7.23) below (cognition acts **bolded**):

(7.22) And now a team led by scientists from the University of Bergen in Norway **thinks** they have an answer. [Space SNA #218]

(7.23) A team of researchers **believes** they have a way to tackle this phenomenon - but the solution is wildly unpopular. [Environment SNA #381]

Both sentences appear early in their respective SNAs and serve to introduce the RA which the writer will report on. Additionally, both clauses position the study as being a solution to a problem. As such, both sentences appear to fulfill similar functions, so the difference in reporting verb may have less to do with author stance and more to do with other concerns, such as avoiding repetition.

Informants explained the use of mental verbs, such as *think* and *believe*, as being engaging for the reader. For certain informants, these verbs represent more closely how people talk in daily life (Informant #1) and help make the SNA more human-centered (Informant #3), which contrasts with the technical, precise, and object-oriented discourse of RAs (Informant #2). As one informant put it, “you don’t really know what other people think,” but “you may know what people have written, what they have said” (Informant #1). Ultimately, this unknowability of what scientists think and believe about their research ties into the subject matter of SNAs:

> *In science journalism, it's more talking about the stuff that we don't know yet or the stuff that is still open for questioning. So I think that's why we tend to use more cerebral words of like, "they think," and "they assume," and "they believe." Do they really know?* [Informant #7]

In other words, since SNAs report on newly published research that has not yet reached a consensus among academics, writers may use more mental verbs to reflect the uncertainty of the findings' implications. These verbs position implications and conclusions as things that are being mulled over in the minds of scientists rather than as objective, cemented fact. Moreover, the less factual nature of mental verbs also shields SNA writers from making claims that a study cannot yet prove (Informant #7), a concern that many SNA writers have (see section 6.4).

As noted above, RA writers used fewer of these verbs overall, but the most frequent included *know* and *assume*. *Know* occurred in the *it+be*+TV+*that*Cl pattern, serving to provide general reference, in nearly all instances (7.24).

(7.24) It is well **known** that tropical forests stand out as the richest biodiverse ecosystems among forest biomes, absorb a vast amount of carbon dioxide and are suppliers of many ecosystem services. [Environment RA #371]

Sentences like (7.24) allow RA writers to position a proposition as a generally accepted fact. This is particularly useful when situating one's argument within an on-going debate. For example, writers of RA #371 wrote excerpt (7.24) after presenting their study's results. Presenting the benefits of maintaining tropical forests as a generally accepted fact increases the argumentative strength of their study as the writers move from presenting the specifics of their research toward situating those findings in a large discussion.

Know in the SNA corpus functioned similarly, though it occurred comparatively often in TV+*that*Cl patterns (7.25).

(7.25) We **know** that one strain of bacteria in our gut, called E. colinissle, can effectively convert ammonia into another safer molecule called arginine. [Health SNA #199]

Thus, the primary differences in the use of *know* between the registers are of grammatical pattern and reliance. That is, *know* was relied upon more heavily by RA writers and largely occurred in the *it+be+*TV+*that*Cl pattern.

Assume, by contrast, was used by RA writers in both *it+be+*TV+*that*Cl patterns and TV+*that*Cl patterns, though largely with non-human subjects. RCs with *assume* showed a range of uses. In some instances, they acted like a more tentative version of *know*, in which the assumed author is a general group of researchers and some knowledge is being attributed to them. In other instances, *assume* RCs served to justify a methodological choice (7.26).

(7.26) Considering the GPS surveys for this year as ground truth, the ASTER areas for this year are accurate to within 1% for the Murray Ice Cap and 3 % for the Simmons Ice Cap. It is **assumed** that this is representative of the accuracy of area mapping from ASTER for the other years. [Environment RA #306]

In instances like (7.26), the assumed author is likely the researchers themselves, as the RC essentially explains the internal thought process behind making the methodological choice. By placing *assume* in the *it+be+*TV+*that*Cl pattern, the writer removes the author from the sentence, in turn conveying a sense that the assumption is more objective than subjective.

7.5 Conclusions

This chapter examined the attribution practices of RA and SNA writers by examining three lexico-grammatical patterns that are often used for attribution, namely sentences with *that* complement clauses (TV+*that*Cl), sentences with an extraposed *that* clause and passive voice verb (*it*+*be*+TV+*that*Cl), and sentences with the prepositional phrase *according to*. SNAs were shown to use more of these patterns of attribution overall, specifically the TV+*that*Cl and *according+to* patterns. RA writers used slightly more *it*+*be*+TV+*that*Cl RCs. These preferences correspond with findings in Chapters 5 and 6, where SNA writers were shown to use more sentences with *that* complement clauses and the RA corpus more sentences with passive verbs.

The prototypical RC used by SNA writers was an initial citation or deictic reference, had a human author, and had a verb that referred to a communicative act or a research act. Attributions often functioned in a chain of reference, where an initial citation introduced a study which would later be referred back to using deictic RCs. By contrast, the prototypical RC used by RA writers had academic-style citation, had a non-human author, and had a verb that referred to a research act. RAs often summarized over a breadth of previous literature using non-integral citation with non-human subjects, which is a more economical method of summarizing over many different previous studies and situating one's research within that literature.

Examination of the reporting verbs used within RCs suggested different patterns of writer and author stance. RA writers relied most on research acts, or verbs referring to the research process, a trend common to science writing in the 'hard' disciplines (Hyland, 2002b). Reliance on these verbs and a corresponding lack of cognition acts produces an empirical and scientific style of discourse. SNA writers, on the other hand, relied more evenly on all three kinds of acts, with discourse act verbs being the most common. Reporting verbs used by SNA writers often suggested either tentativeness or neutrality on the part of the writer. For example, they used certain cognition acts to portray research implications as the product of human reasoning rather than natural logic, as well as certain discourse acts to foreground the researchers' interpretations and background their own. The greater use of cognition acts in SNAs relative to RAs was described by informants as a more natural and entertaining style of news writing.

8 CONCLUSION

In the preceding chapters, I have tried to illustrate some of the ways that one register of popular science writing, the science news article (SNA), compares and contrasts with one register of professional science writing, the research article (RA). This dissertation was designed to compare these two registers. I collected a corpus of 400 SNAs and a corpus of the 400 RAs that the SNAs reported on, allowing me to control for the variable of article topic. As such, the results of linguistic comparisons could be attributed to other situational characteristics, such as who the writer is, who the intended audience is, what the purpose of the register is, and so on. Throughout the chapters, I have also tried to incorporate perspectives from some of the writers who produce SNAs.

I examined three groups of verbal features, each of which was the subject of its own chapter. In Chapter 5, I examined a feature called verb patterns, which refer to the sequences of obligatory elements in independent clauses. I grouped the 30 most frequent verb patterns (e.g., TV+*that*Cl) into 14 more inclusive patterns (e.g., TV+FCl) for easier comparison and analysis. The analysis suggests that RA writing makes greater use of transitive verb patterns with phrasal objects, both in the active and passive voice. SNA writing, by contrast, makes greater use of transitive verb patterns with complement clauses, especially *that* complement clauses. This difference reflects the SNA's purpose to tell stories involving researchers, their words, and their ideas. The RAs rather record observations following strict protocols, often describing those findings and methods in the form of noun phrases. Copular verb patterns were used with similar frequency across the registers, though RAs showed a slight preference for copular patterns with adjective phrase complements and SNAs noun phrase and clausal complements. These patterns were often useful for stating definitions and marking personal stance.

In Chapter 6, I examined variation of the short verb phrase and compared its variation across the registers. Specifically, I measured the rates of finite and non-finite verb phrases, past and present tense verbs, simple and non-simple aspect verbs, active and passive voice verbs, and core and semi- modal verbs. SNAs were shown to be more verbally dense, particularly in terms of finite verb phrases. The comparative lack of verbs in RAs likely reflects the prototypically nominal style of academic prose (Biber & Gray, 2016). The greater presence of methodological description in RAs was also shown in this chapter. RAs showed a preference for the past tense and passive voice, frequent features of RA methods sections (Heslot, 1982). Informants described SNAs as being "present-ist," or focused on the present time to increase news value, and active in style, since the active voice was seen as being more dynamic and interesting to read. SNAs also used many more modal verbs, particularly the possibility modals *can* and *could* and the prediction modals *will* and *might*. Prediction modals often appeared near the beginning of texts, where they served to engage the reader in the content by suggesting hypotheses that could be tested by reading the rest of the article. Moreover, to informants, modal verbs in general serve to protect the writer from making fallacious claims and communicate to readers that individual studies are rarely conclusive on their own.

Chapter 7 examined the use of reporting clauses (RCs) in the registers. While RCs traditionally refer to sentences with a direct report, a speaker, and a verb of communication (e.g., *they said, "…"*), this dissertation examined a wider breadth of RCs, including any attribution of research-related words or ideas packaged in TV+*that*Cl, *it*+*BE*+v+*that*Cl, or *according+to* patterns, since these patterns often serve to report the words and ideas of others (Biber et al., 1999; Charles, 2006). I found that attribution using these three patterns was most frequent in SNAs, particularly via TV+*that*Cl patterns. That SNAs use many TV+*that*Cl patterns relative to

RAs also has implications for the most frequent authors of RCs. SNAs usually attribute words and ideas to human authors (e.g., *the team*), while RAs more often attribute words and ideas to inanimate objects (e.g., *model*). While RAs most often perform attribution through a dense use of non-integral, academic-style citations in order to summarize findings and implications over multiple studies at once, the SNAs analyzed in this study are generally confined to introducing one RA and continually referring back to it throughout the article. When SNA writers do refer to other literature, it is often through general reference in the TV+*that*Cl pattern, where all of humanity (e.g., *we*) or a general academic body (e.g., *astronomers*) is attributed some generally accepted idea. Finally, the reporting verbs used in RCs also suggested differences between the registers. SNA writers use more reporting verbs referring to linguistic acts (e.g., *say*) and mental processes (e.g., *think*) relative to RA writers, who rely largely on verbs referring to parts of the research process (e.g., *calculate*).

8.1 Applications and implications

In the four sections below, I make some final comments on the above conclusions. In section 8.1.1, I discuss what this dissertation can say about the comparison of SNAs and RAs. In section 8.1.2, I discuss the place of SNAs in the context of other registers of popular science. In 8.1.3, I briefly discuss the relevance of verb pattern analysis to other similar work in linguistics. Finally, I discuss potential pedagogical implications resulting from this dissertation in 8.1.4.

8.1.1 On the comparison between SNAs and RAs

SNAs are not simple translations of RAs, though a small number read like brief summaries of them (see section 6.5). Although SNA writers report on recently published RAs, the resulting articles show many differences, which can largely be attributed to three situational characteristics. First, the layout and organization of SNAs is quite different from that of RAs.

RAs are long, detailed texts with informationally dense tables and figures and clearly separated sections fulfilling different rhetorical purposes. SNAs are much briefer, often lack clearly defined sections, and usually exclude informative visuals. These features relate to another important characteristic, namely the purposes of the registers. As one informant explained (see section 4.1.7.5), SNAs function to inform and entertain readers. SNAs are thus shorter and less informationally dense to cater to this entertainment purpose. They linger less on previous literature and use more verbs to counter the nominally dense style of scientific writing. RAs share novel empirical findings or theoretical arguments. They seek to transform knowledge rather than report it (Gotti, 2014). The third interlocking characteristic is the writer. The characteristics of SNA writers are complex. For example, it is untrue that all SNA writers are non-experts. One informant interviewed for this dissertation was both an SNA writer and an active scholar. Another informant had published two non-fiction books on scientific subjects. Others came with strong journalistic backgrounds. In short, there is variation in the expertise and professional experiences of SNA writers; however, despite this variation, these SNA writers do report on research from a consumer's perspective. That is, they report on others' research rather than their own.

While Chapters 5, 6, and 7 attempted to illustrate how these characteristics affected the discourse of RAs and SNAs, it is useful to look at one last example that encapsulates these features. Below, I present short excerpts from the opening paragraph of a space SNA and its matching RA:

> **What would Earth look like to alien astronomers?**
>
> Ever since 1992, when astronomers first discovered two rocky planets orbiting a pulsar in the constellation Virgo, humans have known that other worlds exist beyond our solar system. Today, thanks to the efforts of astronomers and ambitious missions like the now-retired Kepler, we know of more than 4,000 confirmed exoplanets. But if we can see exoplanets orbiting distant stars, that means extraterrestrial observers should be able to see Earth orbiting the Sun. Our tiny blue marble even could be on an alien astronomer's list of rocky exoplanets capable of harboring life. That’s a speculative scenario, of course, but it’s one astronomers still take seriously. In multiple papers over the years, they’ve identified which exoplanets would be able to spot Earth.
>
> Space SNA #287

> **Which stars can see Earth as a transiting exoplanet?**
>
> More than 3000 transiting exoplanets have been detected to date (exoplanets.nasa.gov 2020 July) with dozens of terrestrial planets orbiting in the temperate Habitable Zone of their stars (e.g. Kane et al. 2016; Berger et al. 2018; Johns et al. 2018). NASA’s Transiting Exoplanet Survey Satellite (*TESS*) mission (Ricker et al. 2016) has already searched about 74 per cent of the sky in its 2-yr primary mission for transiting extrasolar planets, including potentially habitable worlds orbiting the closest and brightest stars. The *TESS* Habitable Zone Star Catalog (Kaltenegger et al. 2019) derived the list of stars where *TESS* can detect transiting Earth-sized planets orbiting in the Habitable Zone of their host star.
>
> Space RA #287

The SNA excerpt illustrates several features of the register that we have become familiar with. We find ‘TV+FCl’ and ‘CV+*a*/*an* noun’ verb patterns (*they’ve identified which…, that’s a speculative scenario…*), general references with mental reporting verbs (*humans have known…, we know of…*), a dense use of modal verbs (*can, should, could, would*), and the third-person subject pronoun *they*, highlighting that the article was written by a SNA writer rather than the researcher(s). Other features indicative of the style of SNA writing is certain lexical items (*our tiny blue marble, alien astronomers, we*) and verb contractions (*that's, it's, they've*). Moreover, this brief excerpt covers a lot of ground in its few sentences, ground which resembles familiar moves of popular science discourse, such as orienting the reader to the topic and providing a rationale for the research that is to be reported (Jiang & Qiu, 2022).

The RA excerpt, by contrast, avoids referencing humans in subject positions. There is greater numerical precision, including a more conservative *3,000* exoplanets (as opposed to the SNA article's *4,000*) and the particular percentage of the sky which a satellite has searched in (*about 74 per cent*). Moreover, there is a distinct lack of guidance given to the reader with regard to how to interpret each new proposition. Thus, uninformed readers may be left wondering about the importance of *Kaltenegger et al.*'s *(2019) Catalog* and its derived *list of stars*. As a result, the excerpt and its larger article convey a sense that the text is meant for trained readers like other working scientists.

While there are clear differences between these excerpts and the registers that they represent, it would be unfair to ignore some of their rhetorical and linguistic similarities. For example, the titles of the two excerpts above both ask questions, inviting their audiences to read the article and discover an answer. In addition, while close linguistic analysis would reveal more verbs in the SNA excerpt and more nouns in the RA excerpt, the two excerpts nonetheless would appear more similar than different if they were compared with, say, conversational English, illustrating that both RAs and SNAs are written, informational texts often communicating complex topics. Additional evidence can be found in other literature. Bazerman (1988, Ch. 8) noted how scientists sometimes skim RAs for their major findings and conclusions while skipping methods. Similarly, Berkenkotter and Huckin (1995, Ch. 2) note that some scientists read RAs in a manner analogous to newspaper reading. That is, while RAs are presented in a particular rhetorical order, readers often read in a non-linear pattern looking for newsworthy information. Thus, the greater orientation of SNAs toward highlighting findings and implications and leaving out the background and methodological minutiae is not entirely unfamiliar to even working academics.

In short, this dissertation aimed at providing a unique, generalizable linguistic comparison of popular and professional science writing. It confirmed some earlier arguments about popular science discourse, such as its preference for active rather than passive voice (cf. Seoane & Suárez-Gómez, 2020), added quantitative and qualitative detail to others, such as the use of modal verbs in health-related texts (cf. Csongor and Rébék-Nagy, 2012), and contradicted others, such as the argument that popular science writing avoids hedging to increase newsworthiness (cf. Hyland, 2010). At the same time, it is worth reflecting on similarities between the registers, which, although not the focus of this dissertation, can be gleaned from the preceding chapters.

8.1.2 Contextualizing the SNA in popular science discourse

In section 2.3, I argued that there have been few studies of contemporary science news writing relative to other related registers and genres, especially natively digital ones. Thus, a relevant question to ask in light of this dissertation is what the discourse examined here says about SNA writing and its comparison to those other registers. In short, SNAs suggest certain differences from other forms of popular science, in turn highlighting the need to consider how situational characteristics affect the rhetorical and linguistic features within popular science discourse.

Single-study science news articles, what I have been calling SNAs, are especially influenced by two contextual features, and these features in part distinguish SNAs from other registers of popular science. These features relate to the status of writer and context of publication.

Specifically, despite some outliers, SNA writers are often lay experts rather than active researchers, and SNA writing closely resembles journalism. To illustrate the effect of these characteristics, we can consider a brief excerpt from the research blog *Geneura Team*, which is one of the blogs examined by Luzón (2017). Unlike this dissertation's SNAs, the excerpt below was written by one of the researchers of the studies described in the post. In addition, the writing stems from a blog rather than a news outlet:

> **New research line on µRTS**
>
> This new research line was started one year ago together with PhD student Abdessamed Ouessai and professor Mohammed Salem both from the University of Mascara (Algeria).
>
> The objective is focused on the improvemet *[sic]* of the decision process of an autonomous agent for playing a simple Real-Time Strategy Game, named µRTS (microRTS).
>
> […]
>
> We hope you like it, as usual. :D
>
> (link to blog post)

This excerpt helps to illustrate how SNAs both differ from and are similar to other forms of popular science, such as research blogs. First, the post is written from the first-person perspective, since the writer is also one of the study's researchers (*together with*…, *we hope you*…). In addition, the writer shows their command of scientific English, with the use of abstract nouns (*objective*), nominalizations (*improvement, decision*), and stacked noun modifiers (*the improvement of the decision process of an autonomous agent for playing…*) (see Halliday &

Martin, 1996), suggesting both their experience with professional science writing and a tendency to weave between casual, informal prose and literate, informationally dense prose in blogging. Other differences include the minimally informative title, the lack of an opening ‘hook’, the use of an emoji (*:D*), and the lack of verb patterns used to report others’ words and ideas.

However, it is again worth highlighting some similarities between SNAs and other forms of popular science. For example, the writer of the blog excerpt above uses the phrase *new research* to introduce a study and illustrate its newsworthiness, mentions researcher affiliations (*…both from the University of…*) to credit the researchers and add credibility to the post, and references their mental state (*we hope you like…*) to offer an alternative perspective not typically present in published research writing. When comparing SNAs to more dissimilar contexts, common rhetorical functions can also be found. For example, several of this dissertation’s informants noted the importance of storytelling to connect with SNA readers, a feature which can be seen in the ‘orientation’ move of Three-Minute Thesis talks (Jiang & Qiu, 2022). In addition, Ye’s (2021) science podcasts and Nwogu’s (1991) magazine articles include moves similar to the introductions, methods, results, and discussions of RAs, moves which are also hinted at in this dissertation’s SNA excerpts.

In sum, popular science is a notably messy term (Hilgartner, 1990). It often refers to contexts in which science-themed information is communicated to broad audiences (Gotti, 2014). In the EAP literature, a variety of popular science contexts have been explored, from magazine articles to podcasts to blogs. In this sub-section, I tried to call attention to ways that the language of these contexts differs from and is similar to one another due to contextual characteristics. As a result, paying due attention to these characteristics is paramount for building upon prior research of popular science discourse. For example, it would seem logical to question how the popular science writing of COCA (Liu & Deng, 2015), the BNC (Kaatari, 2019), ICE-GB (Zhang, 2015), ICE-HK (Seoane and Suarez-Gomez, 2020), and the SNA corpus relate to one another within the context of popular science research. Such questions offer intriguing future directions for scholars studying science communication, particularly in corpus-based research.

8.1.3 Corpus-based descriptions of verb patterns

Arguably, one of the more unique contributions of this dissertation is its linguistic description of verb patterns in English. Concepts similar to verb patterns have been investigated elsewhere under different names and for different purposes. Ellis, Römer, and O'Donnell (2016) adopted a construction grammar approach to their description of Verb Argument Constructions (VACs) in English texts drawn from the BNC. VACs are patterns of verbs and usually some other element(s) (e.g., verb *across* noun) that develop a particular form-meaning pairing. However, in comparison to the current dissertation, the VAC approach usually adopts a second language acquisition (SLA) aim (e.g., Ma & Qian, 2020) and examines fine-grained patterns (e.g. v *with* n, v *for* n, v *into* n) whose descriptions may prove useful for understanding English acquisition and teaching.

In addition to VAC research, there are other large-scale projects investigating similar patterns. For example, Francis et al.'s (1996) book on pattern grammar identified many of the patterns that would later be investigated in the form of VACs. Similarly, both Allerton (1982) and Herbst et al. (2004) provide in-depth discussion and description of English valency patterns. However, Francis et al. did not provide frequency counts for their patterns, Allerton's work is not corpus-based and thus does not provide frequency information, and Herbst et al. (2004) provides only general frequency information for the valency patterns of nouns, verbs, and adjectives (e.g., 'n *advised as* n' is 'rare' in their corpus). Finally, Biber et al.'s (1999) corpus-based grammar of

English, while clearly influential on this dissertation, only provides partial information on the frequency of different valency patterns, such as the proportional frequency of certain patterns used by certain verbs (e.g., *consider* prefers the TV+PhO pattern; p. 387).

Thus, the current dissertation, specifically Chapter 5, contributes in a small way to this literature by providing an empirical comparison of verb patterns across two English language registers for the purposes of discourse analysis. As a result, this dissertation can illustrate that transitive verbs with phrasal objects are likely the most frequent in certain literate registers such as academic writing and news reportage. Other patterns, such as transitive verbs with finite complement clauses, are also common but will vary according to whether reporting others' words and ideas is an important function of that register. As verb patterns have not been routinely studied in the discourse analysis of registers and genres, there is an opportunity to fill this gap given the increasing availability and sophistication of computer tools.

8.1.4 Pedagogical implications of popular science discourse in EAP

In this section, I reflect on what pedagogical implications arise from this study of RA and SNA writing. Below, I layout the arguments for adopting popular science discourse for teaching and learning and then discuss how these arguments apply to the discourse of the SNA.

One argument for the use of popular discourse in teaching science reading, writing, or content is that its linguistic profile lends itself to easier reading by non-experts, such as students reading in an additional language. For example, studies of cohesion find that popular science texts may aid reading comprehension by using a simpler and more consistent thematic structure (Nwogu & Bloor, 1991), as well as lexical cohesive devices that aid the non-specialist in understanding science (Myers, 1991). Similarly, some texts make a greater use of lists and linking adverbials to organize discourse and guide the non-expert reader (Hyland & Zou, 2020).

Other strategies that writers engage in to calibrate to their general audience include the use of paraphrases, metaphors, rhetorical questions, and humor (Luzón, 2013). Moreover, register studies indicate that some popular science texts mark a mid-way point between academic prose and newspaper writing in terms of grammatical complexity (Larsson & Kaatari, 2020), stance marking (Larsson, 2019), and technical lexis (Padula et al. 2020).

Another argument is that popular discourse may offer a useful conceptual tool for better understanding the discourse of professional science. Hyland (2010) argues that, although popular science presents a different model of science than research articles, such differences may have an attention-raising function for students, and as such can raise their awareness to the rhetorical and linguistic styles of popular and professional science discourse. Parkinson and Adendorff (2004) come to a similar conclusion, saying that, "besides usually being easier in terms of content and thus making for a suitable first reading/introduction to a topic […] popular texts may be valuable both in the teaching of science and in the teaching of scientific writing" (p. 391-392). The authors then provide several ways in which popular science texts might be valuable, including that they allow students more personality in scientific writing, that they introduce students to debated scientific knowledge before it becomes canonized in textbooks, and that they may help students become more aware of a fuller range of scientific genres.

Pedagogical studies have pointed toward the validity of the above arguments. For example, Parkinson's (2000) use of science magazine articles and non-fiction books in a science content-based English language course in South Africa points to the usefulness of popular science texts in content-based classrooms. A number of other studies have investigated various aspects of popular science writing for STEM majors in native and non-native English contexts. For example, introducing undergraduate STEM majors in Sweden to popular science writing changed their perspective on their research (Pelger & Nilsson, 2016). Students reported positive attitudes toward writing popular articles and gleaned new perspectives on science writing, such as identifying the "big picture" of their work (Pelger, 2017). This sense of developing audience awareness was also reflected by Crone et al.'s (2011) undergraduate STEM students in the U.S. Moreover, learning to write in the style of popular science may not negatively affect students' skills in (non-popular) academic writing (Rakedzon & Baram-Tsabari, 2016).

These studies propose intriguing characteristics of popular science writing and its application for teaching and learning the language of science communication. However, as cautioned in section 8.1.2, all registers of popular science are not the same and thus cannot be considered equally useful for pedagogical purposes. Thus, a question arises whether SNAs can be useful for such purposes. Certainly, this dissertation did not investigate the entire range of features cited above and therefore cannot speak to the cohesion, grammatical complexity, use of lists, metaphor, and academic vocabulary of SNAs. However, for some of these features, implications can be noted. For example, informants' insistence that SNAs be tailored with the general reader in mind suggests that these texts are suitable entry points to topics for which textbooks or research articles may not be. This idea is supported by linguistic evidence, such as the heavy use of (the less grammatical complex) *that* complement clauses (see Biber et al., 2011) and the comparatively verbal style of SNAs. Moreover, the relatively frequent use of direct quotations also suggests that these texts present a linguistic style more concordant with non-scientific discourse (cf. Halliday & Martin, 1996).

However, as others have noted, a popular register like the SNA is not an accurate model of a professional register like the RA (Hyland, 2010; Parkinson & Adendorf, 2004). Thus, one of the greatest strengths of popular science writing for teaching and learning is not as an easier-to-read version of research writing but rather as a different perspective on scientific research and scientific communication. For example, in section 7.2.1, I compared the use of certain interview excerpts in SNAs with Gilbert and Mulkays' (1984) notion of the 'contingent repertoire,' which contrasts with the type of reasoning and linguistic style found in published research writing. Thus, in some cases, when students read an SNA including interview excerpts, they are not simply reading a spoken version of the research process but rather a different perspective on it entirely. Moreover, as section 5.4.1 showed, not all SNAs are the same, particularly as it relates to the amount of space given to direct and indirect reports. In certain circumstances, such as content-based classes in which the scientific content of the article is most important, SNAs in which the writer relies less on quotations and more on their own interpretations and explanations may be more useful.

In short, SNAs present some of the features that make popular science discourse more accessible by non-expert readers. Moreover, while some SNAs do read like brief summaries of the longer RAs (see section 6.5), many do not, and as such they present insightful comparisons with professional science discourse that can lead to a better understanding of both registers. However, SNAs also present some weaknesses. SNAs adopt a notably journalistic style, particularly in their use of direct quotations with reporting clause-style syntax. Such texts may be less desirable in EAP classrooms, particularly if the topic is scientific content or scientific style.

8.2 Limitations

It is worth noting certain limitations of this dissertation. Below I focus on two limitations relating to the corpora used and the (in)completeness of the verb pattern analysis.

The first limitation relates to the RA corpus compiled for this dissertation. While the atypical inclusion criteria have already been noted (see section 3.1.1), I want to mention a different limitation here, namely the fact that these RAs were not separated by their constituent sections. The ways that the individual parts of RAs, such as the introduction, methods, and so on, fulfill different communicative functions are important considerations for this genre. For example, Swales (2004) argues that the methods section is a crucial location of disciplinary differences, with some RAs having extensive methods, others limited methods, and some no methods at all. RA parts also adopt different linguistic devices to achieve their rhetorical purposes (Heslot, 1982). Thus, analyzing this dissertation's RAs by section would have added additional insight into how SNAs compare and contrast with different parts of the RA. As another example, the distribution of modal verbs (see section 6.4) in RAs likely varies as a function of part-RA, and this added detail would help better compare SNAs with RAs. The primary reason for not analyzing the RAs by section is that the RAs were not selected on the basis of their macro-structures but instead by their reference in SNAs. The current RA corpus would need to be significantly changed to have 400 texts with the same organization.

In addition, the analysis of verb patterns in Chapter 5 would benefit from improvements to the program and a greater focus on more patterns. As noted in 3.1.4.1, intransitive verbs were particularly difficult to automatically identify, resulting in the need for a small dictionary of common intransitive verbs in order for the program to return counts of these patterns. While this dictionary still provides strong coverage of the most common intransitive verbs in the corpora, it is incomplete and thus cannot reflect the true rate of intransitive verb patterns in the registers. Additionally, I confined my analysis only to those verb patterns of the main clause. As a result, I did not provide any counts or analysis of the patterns of finite or non-finite dependent clauses, such as adverbial clauses, relative clauses, and noun complement clauses. Given that I earlier argued that this dissertation adds to the relative dearth of statistical information on the verb patterns of different registers (see section 8.2.3), the incompleteness of the information provided in Chapter 5 should then be measured alongside its contributions.

8.3 Areas of future research

Section 8.1 hinted at certain areas of future research. Sections 8.1.1 and 8.1.2 focused on the comparison of SNAs with professional science writing and with other registers of popular science writing. In my view, there is room for investigating, either from a linguistic or rhetorical perspective, variation within popular science discourse. To date, studies of the expanding contexts of science communication, many of which could be described as contexts of popular science, often examine these contexts one at a time, while many of the comparative studies, like this dissertation, focus on comparing popularizations with RAs. One exception is Fu and Hyland (2014), who compared certain rhetorical features of science magazine articles and opinion pieces published in newspapers. Studies like this one shift the focus from, *how do writers popularize research writing?* to, *how do the characteristics of different contexts of popular science affect the features of their texts?* One intriguing comparison is how the language of researchers who popularize science compares with the language of non-researchers who popularize science, such as science journalists or university communication specialists. A brief comparison was explored in section 8.1.2 (see also Myers, 1999, Ch. 5). Comparisons could be made between many contexts at once, an approach which would be well suited for register analysis, especially Multidimensional Analysis (MDA). For example, an MDA study could sketch in detail the linguistic profiles of those contexts illustrated in Hilgartner's (1990) stream of scientific communication (see Figure 2.1). Findings may challenge or confirm commonly held beliefs about what genres or registers are more 'popular' and which are more 'professional'.

Section 8.1.3 highlighted some arguments for the use of popular texts in classrooms, such as EAP or English for Science and Technology (EST) classrooms. However, many of these arguments are based on textual analysis, and classroom-based studies are largely from a STEM education perspective rather than an EAP or EST one. This dissertation only furthers this discord by providing possible pedagogical implications without directly examining what those implications look like in classrooms. Parkinson’s (2000) description of an EST classroom in South Africa is a good example of a study that does examine the use of popular science discourse in the classroom, but there are few others. One research strain worth exploring is the extent to which popular science discourse is already in use in various ESL or EFL settings. For example, personal anecdotal experience suggests that some popular texts have an intuitive appeal for institutions that prepare students for, or test their ability to use, university-level English. Muñoz (2015) also explains that certain university contexts in Argentina make frequent use of (semi-)popularization articles for the reasons outlined in 8.1.4. With a solid understanding of who uses what popular science discourse and for what purposes, pedagogically-minded studies could develop research questions that are most relevant for improving the use of these texts.

REFERENCES

Adams-Smith, D. E. (1987). The process of popularization—rewriting medical research papers for the layman: Discussion paper. *Journal of the Royal Society of Medicine, 80,* 634-636.

Agresti, A. (2007). *An introduction to categorical data analysis*. New Jersey: Wiley.

Allerton, D. J. (1982). *Valency and the English verb*. London: Academic Press.

Anthony, L. (2018). *Introducing English for Specific Purposes*. London & New York: Routledge.

Barton, R. (2003). 'Men of science': Language, identity and professionalization in the mid-Victorian scientific community. *History of Science, 41*(1), 73-119.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Bauer, M. (1998). The medicalization of science news – from the "rocket-scalpel" to the "gene-meteorite complex." *Social Science Information, 37*(4), 731–751.

Bauer, M., & Bucchi, M. (Eds.) (2007). *Journalism, science, and society: Science communication between news and public relations*. London & New York: Routledge.

Bauer, M., Durant, J., Ragnarsdottir, A., & Rudolphsdottir, A. (1995). *Science and technology in the British press, 1946-1990*. London: The Science Museum.

Bawarshi, A., & Reiff, M. J. (2010). *Genre: An introduction to history, theory, research and pedagogy*. West Lafayette & Fort Collins: Parlor Press and WAC Clearinghouse.

Bazerman, C. (1988). *Shaping written knowledge*. Madison: The University of Wisconsin Press.

Belcher, D. D. (2013). The future of ESP research: Resources for access and choice. In B. Paltridge & S. Starfield (Eds.), *Handbook of English for Specific Purposes,* (pp. 535-551). West Sussex: Wiley-Blackwell.

Belcher, D. D. (2023). Digital genres: What they are, what they do, and why we need to better understand them. *English for Specific Purposes*, 70, 33-43.

Bensaude-Vincent, B. (2001). A genealogy of the increasing gap between science and the public. *Public Understanding of Science, 10,* 99-113.

Berkenkotter, C., & Huckin, T. (1995). Genre knowledge in disciplinary communication: Cognition/culture/power. London & New York: Routledge.

Biber, D., & Conrad, S. (2019). *Register, genre, and style.* (2nd Ed.). Cambridge: Cambridge University Press.

Biber, D., & Gray, B. (2013). Nominalizing the verb phrase in academic science writing. In B. Aarts, J. Close, G. Leech, & S. Wallis (Eds.), *The verb phrase in English,* (pp. 99-132). Cambridge: Cambridge University Press.

Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing.* Cambridge: Cambridge University Press.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5-35.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow: Pearson Education Limited.

Blanchard, A. (2011). Science blogs in research and popularization of science: Why, how and for whom? In M. Cockell, J. Billotte, F. Darbellay & F. Waldvogel (Eds.), *Common knowledge: The challenge of transdisciplinarity* (pp. 219-232). Lausanne, Switzerland: EPFL Press.

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide for beginners*. London: Sage.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.

Bucchi, M., & Trench, B. (Eds.) (2014). *Routledge handbook of public communication of science and technology*. London & New York: Routledge.

Buehl, J. (2022). Graphical abstracts: Visually circulating scientific arguments. In C. Hanganu-Bresch, M. J. Zerbe, G. Cutrufello, & S. M. Maci (Eds.), *The Routledge handbook of scientific communication* (pp. 290-306). London & New York: Routledge.

Calsamiglia, H. (2003). Popularization discourse. *Discourse Studies, 5*(2), 139-146.

Calsamiglia, H., & Ferrero, C.L. (2003). Role and position of scientific voices: Reported speech in the media. *Discourse Studies, 5*(2), 147-173.

Calsamiglia, H., & Van Dijk, T. (2004). Popularization discourse and knowledge about the genome. *Discourse & Society, 15*(4), 369-389.

Casal, J., Shirai, Y., & Lu, X. (2022). English verb-argument construction profiles in a specialized academic corpus: Variation by genre and discipline. *English for Specific Purposes, 66,* 94-107.

Charles, M. (2006). Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines. *English for Specific Purposes, 25,* 310-331.

Clarke, V. (2006). 'Gay men, gay men and more gay men': traditional, liberal and critical perspectives on male role models in lesbian families. *Lesbian & Gay Psychology Review, 7,* 19–35.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd Edition). Hillsdale: Lawrence Earlbaum Associates.

Conrad, S. (1996). *Academic discourse in two disciplines: professional writing and student development in biology and history*. PhD dissertation, Northern Arizona University.

Conrad, S. (2014). Expanding multi-dimensional analysis with qualitative research techniques. In T. B. Sardinha & M. V. Pinto (Eds.), *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber*, (pp. 273-295). Amsterdam & Philadelphia: John Benjamins.

Crone, W., Dunwoody, S., Rediske, R., Ackerman, S., Zenner Peterson, G., & Yaros, R. (2011). Informal science education: A practicum for graduate students. *Innov High Educ, 36*, 291-304.

Csongor, A, & Rébék-Nagy, G. (2012). Hedging in popular scientific articles on medicine. *Acta Medica Marisiensis, 59*(2), 98-99.

Cumming, G., Fidler, F., & Vaux, D. L. (2007). Error bars in experimental biology. *The Journal of Cell Biology, 177*(1), 7-11.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research, 28*(3), 369-382.

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,* 357-366.

Delucchi, K. L. (1993). On the use and misuse of chi-square. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences* (pp. 294-319). Hillsdale: Lawrence Earlbaum Associates.

Ding, D. D. (2002). The passive voice and social values in science. *Journal of Technical Writing and Communication, 32*(2), 137-154.

Dolnicar, S., Grün, B., Leisch, F., & Schmidt, K. (2014). Required sample sizes for data-driven market segmentation analyses in tourism. *Journal of Travel Research, 53*(3), 296-306.

Dunwoody, S. (2014). Science journalism in the digital age. In M. Bucchi & B. Trench (Eds.), *Routledge Handbook of Public Communication of Science and Technology* (pp. 15-26). London & New York: Routledge.

Egbert, J., Biber, D., & Gray, B. (2022). *Designing and evaluating language corpora: A practical framework for corpus representation*. Cambridge: Cambridge University Press.

Einsiedel, E. F. (1992). Framing science and technology in the Canadian press. *Public Understanding of Science*, *1*(1), 89–101.

Ellis, N., Römer, U., O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar*. West Sussex: Wiley-Blackwell.

Facts & Figures (n.d.). Retrieved from https://spacy.io/usage/facts-figures#benchmarks.

Fahnestock, J. (1986). Accommodating science: The rhetorical life of scientific facts. *Written Communication, 3*(3), 275-296.

Fontenelle, B. (1990). *Conversations on the Plurality of Worlds.* (H.A. Hargreaves, trans.). Berkeley: University of California Press. (Original work published 1686).

Fox, J. & Weisberg, S. (2019). *An {R} Companion to Applied Regression* (3rd Ed.). Thousand Oaks: Sage.

Francis, G., Hunston, S., & Manning, E. (1996). *Collins COBUILD grammar patterns 1: Verbs*. London: Harper Collins.

Fu, X., & Hyland, K. (2014). Interaction in two journalistic genres: A study of interactional metadiscourse. *English Text Construction, 7*(1), 122-144.

Garcés-Conejos, P. & Sánchez-Macarro, A. (1998). Scientific discourse as interaction: Scientific articles vs. popularizations. In A. Sánchez-Macarro & R. Carter. (Eds.). *Linguistic choices across genres*, (pp. 173-190). Amsterdam & Philadelphia: John Benjamins.

Garnier, M., & Schmitt, N. (2014). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research, 19*(6), 646-666.

Giannoni, D. S. (2008). Popularizing features in English journal editorials. *English for Specific Purposes, 27*, 212-232.

Gilbert, G. N., & Mulkay, M. (1984). *Opening Pandora's Box: A Sociological Analysis of Scientific Discourse*. Cambridge: Cambridge University Press.

Glesne, C. (2014). *Becoming qualitative researchers: An introduction.* (5th Ed.). Boston: Pearson.

Gotti, M. (2014). Reformulation and recontextualization in popularization discourse. *Iberica, 27,* 15-34.

Gray, B. (2015). *Linguistic variation in research articles: When discipline tells only part of the story* (Studies in Corpus Linguistics, Vol. 71). Amsterdam & Philadelphia: John Benjamins.

Grundmann, R., & Cavaillé, J-P. (2000). Simplicity in science and its publics. *Science as Culture, 9,* 353-389.

Halliday, M. A. K., & Martin, J. R. (1996). *Writing Science*. Bristol: The Falmer Press.

Halliday, M. A. K., & Matthiessen, C. (2004). *An introduction to functional grammar* (3rd Ed.). London: Hodder Arnold.

Hashemi, H. (2012). Reflections on mixing methods in applied linguistics research. *Applied Linguistics, 33*(2), 206-212.

Hashemi, H., & Babaii, E. (2012). Exploring the nature of mixing methods in ESP research. *ESP Across Cultures, 9,* 115-134.

Herbst, T., Heath, D., Roe, I. F., & Götz, D. (2004). *A valency dictionary of English: A corpus-based analysis of the complementation patterns of English verbs, nouns, and adjectives*. Berlin & New York: De Gruyter.

Heslot, J. (1982). Tense and other indexical markers in the typology of scientific texts in English. In J. Hoedt, L. Lundquist, H. Picht, & J. Quistgaard (Eds.), *Pragmatics and LSP* (pp. 83-103). Copenhagen: Copenhagen School of Economics.

Hewings, M., & Hewings, A. (2002). "It is interesting to note that…": A comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes, 21,* 367-383.

Hilgartner, S. (1990). The dominant view of popularization: Conceptual problems, political uses. *Social Studies of Science, 20*(3), 519-539.

Hu, G., & Liu, Y. (2018). Three minute thesis presentations as an academic genre: A cross-disciplinary study of genre moves. *Journal of English for Academic Purposes, 35,* 16-30.

Huddleston, R. (1971). *The sentence in written English: a syntactic study based on an analysis of scientific texts*. Cambridge, Cambridge University Press.

Hyland, K. (2002a). Specificity revisited: How far should we go now? *English for Specific Purposes, 21*(4), 385–395.

Hyland, K. (2002b). Activity and evaluation: Reporting practices in academic writing. In. J. Flowerdew (Ed.), *Academic discourse*, (pp. 115-130). London: Longman.

Hyland, K. (2007). *Disciplinary discourses: Social interactions in academic writing*. 2nd edition. Ann Arbor: University of Michigan Press.

Hyland, K. (2010). Constructing proximity: Relating to readings in popular and professional science. *Journal of English for Academic Purposes*, 9, 116-127.

Hyland, K. (2022). The scholarly publishing landscape. In C. Hanganu-Bresch, M. J. Zerbe, G. Cutrufello, & S. M. Maci (Eds.), *The Routledge Handbook of Scientific Communication* (pp. 15-25). London & New York: Routledge.

Hyland, K., & Zou, H. (2020). In the frame: Signalling structure in academic articles and blogs. *Journal of Pragmatics*, *165*, 31-44.

Hyland, K., & Jiang, F. (2019). *Academic discourse and global publishing: Disciplinary persuasion in changing times*. London & New York: Routledge.

Jiang, F., & Qiu, X. (2022). Communicating disciplinary knowledge to a wide audience in 3MT presentations: How students engage with popularization of science. *Discourse Studies, 24*(1), 115-134.

Lan, G., Liu, Q., & Staples, S. (2019). Grammatical complexity: ‘What does it mean’ and ‘So What’ for L2 writing classrooms?

Larsson, T. (2019). Grammatical stance marking across registers. *Register Studies*, *1*(2), 243-268.

Larsson, T., & Kaatari, H. (2020). Syntactic complexity across registers: Investigating (in)formality in second-language writing. *Journal of English for Academic Purposes, 45,* 1-16.

Leech, G. (2004). *Meaning and the English verb*. London & New York: Longman.

Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in contemporary English: A grammatical study*. Cambridge: Cambridge University Press.

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam & Philadelphia: John Benjamins.

Lightman, B. (2007). *Victorian Popularizers of Science: Designing Nature for New Audiences.* Chicago: University of Chicago Press.

Lillis, T., & Curry, M. J. (2010). *Academic writing in a global context*. London & New York: Routledge.

Liu, C.-Y. (2023). Suitability of TED-Ed animations for academic listening. *English for Specific Purposes, 72,* 4-15.

Liu, Q., & Deng, L. (2017). A genre-based study of shell-noun use in the N-be-that construction in popular and professional science articles. *English for Specific Purposes, 48,* 32-43.

Liu, Y., Tang, R., & Lim, F. V. (2023). The use of code glosses in three minute thesis presentations: A comprehensibility strategy. *Journal of English for Academic Purposes,* 1-14.

Luzón, M. J. (2013). Public communication of science in blogs: Recontextualizing scientific discourse for a diversified audience. *Written Communication, 30*(4), 428-457.

Luzón, M. J. (2017). Connecting genres and languages in online scholarly communication: An analysis of research group blogs. *Written Communication, 34*(4), 441-471.

Luzón, M. J. (2019). Bridging the gap between experts and publics: The role of multimodality in disseminating research in online videos. *Iberica, 37,* 167-192.

Luzón, M. J. (2023). Multimodal practices of research groups in Twitter: An analysis of stance and engagement. *English for Specific Purposes, 70,* 17-32.

Luzón, M. J., & Pérez-Llantada, C. (2019) (Eds.). *Science communication on the Internet: Old genres meet new genres.* Amsterdam & Philadelphia: John Benjamins.

Ma, H., & Qian, M. (2020). The creation and evaluation of a grammar pattern list for the most frequent academic verbs. *English for Specific Purposes, 58*, 155-169.

Mann, S. (2011). A critical review of qualitative inter- views in applied linguistics. *Applied Linguistics*, *32*, 6-24.

Master, P. (1991). Active verbs with inanimate subjects in scientific prose. *English for Specific Purposes, 10,* 15-33.

Mehlenbacher, A. (2017). Crowdfunding science: Exigencies and strategies in an emerging genre of science communication. *Technical Communication Quarterly, 2*6(2), 127-144.

Miller, N., Budgell, B., & Fuller, K. (2012). 'Use the active voice whenever possible': The impact of style guidelines in medical journals. *Applied Linguistics, 34*(4), 393-414.

Moisl, H. (2015). *Cluster analysis for corpus linguistics*. Berlin & New York: De Gruyter.

Moriarty, D., & Mehlenbacher, A. (2019). The coaxing architecture of Reddit's r/science: Adopting *ethos*-assessment heuristics to evaluate science experts on the Internet. *Social epistemology, 33*(6), 514-524.

Muñoz, V. L. (2015). The vocabulary of agriculture semi-popularization articles in English: A corpus-based study. *English for Specific Purposes*, *39*, 26-44.

Myers, G. (1989). The pragmatics of politeness in scientific articles. *Applied Linguistics, 10*(1), 1-35.

Myers, G. (1991). Lexical cohesion and specialized knowledge in science and popular science texts. *Discourse Processes, 14*, 1-26.

Myers, G. (1999). Writing biology: Texts in the social construction of scientific knowledge. Madison: The University of Wisconsin Press.

Myers, G. (2003). Discourse studies of scientific popularization: Questioning the boundaries. *Discourse Studies*, *5*, 265–279.

Nakagawa, S., Johnso, Paul C. D., & Schielzeth, H. (2020). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J. R. Soc. Interface*, *14*, 1-11.

Nwogu, K. N. (1991). Structure of science popularizations: A genre-analysis approach to the schema of popularized medical texts. *English for Specific Purposes, 10,* 111-123.

Nwogu, K. N., & Bloor, T. (1991). Thematic progression in professional and popular medical texts. In E. Ventola (Ed.), *Functional and Systemic Linguistics: Approaches and Uses* (pp. 369-384). Berlin & New York: De Gruyter.

Padula, M., Panza, C., Muñoz, V. (2020). The pronoun *this* as a cohesive encapsulator in engineering semi-popularization articles written in English. *Journal of English for Academic Purposes, 44,* 1-10.

Palmer, F.R. (1990). *Modality and the English modals*. (2nd Ed.) London & New York: Routledge.

Partridge, B. & Starfield, S. (Eds.) (2013). *The handbook of English for specific purposes*. West Sussex: Wiley-Blackwell.

Parkinson, J. (2000). Acquiring scientific literacy through content and genre. *English for Specific Purposes, 19,* 369-387.

Parkinson, J., & Adendorff, R. (2004). The use of popular science articles in teaching scientific literacy. *English for Specific Purposes*, *23*, 379-396.

Patton, M. Q., (2015). *Qualitative research and evaluation methods*. (4th Ed). London: Sage.

Pelger, S., (2017). Popular science writing brings new perspectives into science students' theses. *International Journal of Science Education, Part B,* 1-17.

Pelger, S., & Nilsson, P. (2016). Popular science writing to support students' learning of science and scientific literacy. *Research in Science Education, 46,* 439-456.

Pellechia, M.G. (1997). Trends in science coverage: A content analysis of three US newspapers. *Public Understanding of Science, 6*(1), 49–68.

Pérez-Llantada, C. (2016). How is the digital medium shaping research genres? Some cross-disciplinary trends. *ESP Today, 4*(1), 22-42.

Pérez-Llantada, C. (2021). Grammar features and discourse style in digital genres: The case of science-focused crowdfunding projects. *Revista Signos: Estudios de Lingüística, 54*(105), 73-96.

Python Software Foundation. (2021). Python Language Reference, version 3.10. Available at http://www.python.org.

Qiu, X., & Jiang, F. (2021). Stance and engagement in 3MT presentations: How students communicate disciplinary knowledge to a wide audience. *Journal of English for Academic Purposes, 51*, 1-12.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language.* London & New York: Longman.

R Core Team (2020). *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rakedzon, T., & Baram-Tsabari, A. (2016). Assessing and improving L2 graduate students' popular science and academic writing in an academic writing course. *Educational Psychology*, 1-19.

Salager-Meyer, F. (1992). A text-type and move analysis study of verb tense and modality distribution in medical English abstracts. *English for Specific Purposes, 11*, 93-113.

Samraj, B. (2016). Research articles. In K. Hyland & P. Shaw (Eds.), *The Routledge Handbook of English for Academic Purposes*. pp. 403-415. London & New York: Routledge.

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, *55*(3), 182-186.

Seoane, E., & Suárez-Gómez, C. (2020). Elaboration, compression and explicitness across sub-registers of popular and academic writing in Hong Kong English. *Register Studies, 2*(2), 275-305.

Sharpe, D. (2015). Your Chi-square test is statistically significant: Now what? *Practical Assessment, Research & Evaluation, 20*(8). 1-10.

Sheskin, D. J. (2011). Handbook of parametric and nonparametric statistical procedures. (5th ed.). *New York: CRC Press*.

Sidler, M. (2016). "The Chemistry Liveblogging Event: The Web Refigures Peer Review." In A. G. Gross & J. Buehl (Eds.), *Science and the Internet: Communicating Knowledge in a Digital Age*, (pp. 99–116). London & New York: Routledge.

spaCy: Industrial-strength Natural Language Processing in Python. (2022). Retrieved from https://spacy.io/.

Swales, J. (1981). *Aspects of article introductions*. Language Studies Unit, University of Aston in Birmingham.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.

Swales, J., & Feak, C. (2012). *Academic Writing for Graduate Students: Essential Tasks and Skills*, 3rd Ed. Ann Arbor: University of Michigan Press.

Tarone, E., Dwyer, S., Gillette, S., & Icke, V., (1998). On the use of the passive and active voice in astrophysics journal papers: With extensions to other language and other fields. *English for Specific Purposes, 17*(1), 113-132.

Thompson, G., & Ye, Y. (1991). Evaluation in the reporting verbs used in academic papers. *Applied Linguistics, 12*(4), 365–382.

Trench, B. (2008). Internet: Turning science communication inside-out? In M. Bucchi & B. Trench (Eds.), *Handbook of Public Communication of Science and Technology*, (pp. 186-198), London & New York: Routledge.

Trimble, L. (1985). *English for science and technology: A discourse approach.* Cambridge: Cambridge Language Teaching Library.

Varttala, T. (1999). Remarks on the communicative functions of hedging in popular scientific and specialist research articles on medicine. *English for Specific Purposes, 18*(2), 177-200.

Valeiras-Jurado, J., & Bernad-Mechó, E. (2022). Modal density and coherence in science dissemination: Orchestrating multimodal ensembles in online TED talks and youtube science videos. *Journal of English for Academic Purposes, 58,* 1-12.

Van Dijk, T. (1988). *News as discourse*. Hillsdale: Lawrence Erlbaum.

Wallis, S. (2021). *Statistics in corpus linguistics research: A new approach*. London & New York: Routledge.

Wells, R. (1960). Nominal and verbal style. In T. A. Sebeok (Ed.), *Style in Language*, (pp. 213-220). Cambridge: MIT Press.

Whitley, R. (1985). Knowledge producers and knowledge acquirers: Popularisation as a relation between scientific fields and their publics. In. T. Shinn & R. Whitley (Eds.), *Expository science: Forms and functions of popularisation.* (pp. 3-28). Dordrecht: D. Reidel Publishing Company.

Wickham, H. (2021). *rvest*: Easily Harvest (Scrape) Web Pages. R package version 1.0.0. https://CRAN.R-project.org/package=rvest

Wingrove, P. (2017). How suitable are TED Talks for academic listening? *Journal of English for Academic Purposes, 30*, 79-95.

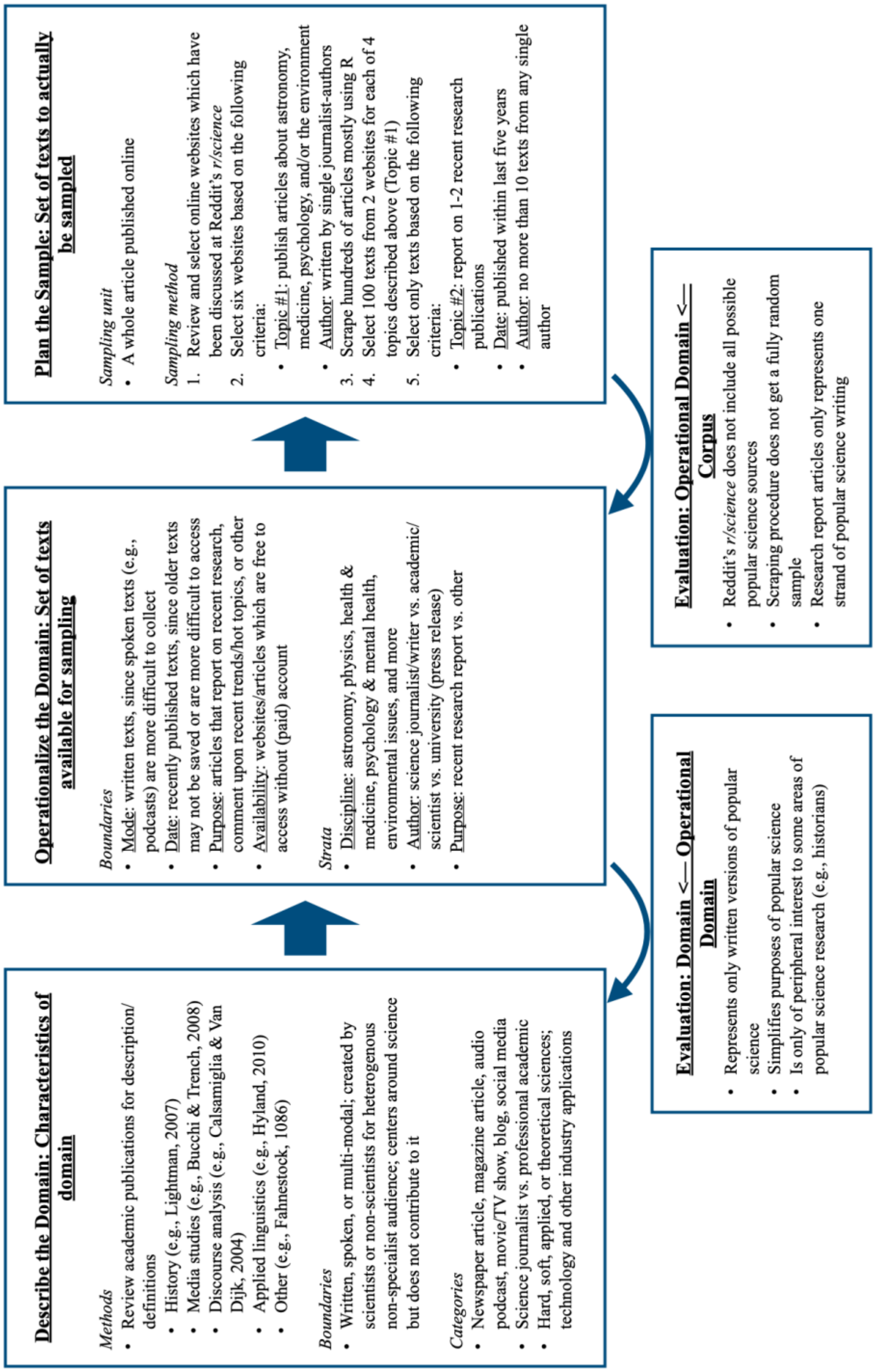
Xia, S. (2023). Transcending science in scientific communication: Multimodal strategies to incorporate humanistic perspectives in TED talks on biology. *English for Specific Purposes, 71*, 60-77.

Ye, Y. (2021). From abstracts to “60-second science” podcasts: Reformulation of scientific discourse. *Journal of English for Academic Purposes, 53*, 1-13.

Zhang, G.B. (2015). It is suggested that…or it is better to…? Forms and meanings of subject it-extraposition in academic and popular writing. *Journal of English for Academic Purposes, 20*, 1-13.

APPENDICES

Appendix A: Domain considerations for the SNA corpus (adapted from Egbert, Biber, and Gray (2022))



Describe the Domain: Characteristics of domain

Methods

- Review academic publications for description/definitions
  - History (e.g., Lightman, 2007)
  - Media studies (e.g., Bucchi & Trench, 2008)
  - Discourse analysis (e.g., Calsamiglia & Van Dijk, 2004)
  - Applied linguistics (e.g., Hyland, 2010)
  - Other (e.g., Fahnestock, 1086)

Boundaries

- Written, spoken, or multi-modal; created by scientists or non-scientists for heterogenous non-specialist audience; centers around science but does not contribute to it

Categories

- Newspaper article, magazine article, audio podcast, movie/TV show, blog, social media
- Science journalist vs. professional academic
- Hard, soft, applied, or theoretical sciences; technology and other industry applications

Operationalize the Domain: Set of texts available for sampling

Boundaries

- Mode: written texts, since spoken texts (e.g., podcasts) are more difficult to collect
- Date: recently published texts, since older texts may not be saved or are more difficult to access
- Purpose: articles that report on recent research, comment upon recent trends/hot topics, or other
- Availability: websites/articles which are free to access without (paid) account

Strata

- Discipline: astronomy, physics, health & medicine, psychology & mental health, environmental issues, and more
- Author: science journalist/writer vs. academic/scientist vs. university (press release)
- Purpose: recent research report vs. other

Evaluation: Domain <— Operational Domain

- Represents only written versions of popular science
- Simplifies purposes of popular science
- Is only of peripheral interest to some areas of popular science research (e.g., historians)

Plan the Sample: Set of texts to actually be sampled

Sampling unit

- A whole article published online

Sampling method

1. Review and select online websites which have been discussed at Reddit's *r/science*
2. Select six websites based on the following criteria:
   - Topic #1: publish articles about astronomy, medicine, psychology, and/or the environment
   - Author: written by single journalist-authors
3. Scrape hundreds of articles mostly using R
4. Select 100 texts from 2 websites for each of 4 topics described above (Topic #1)
5. Select only texts based on the following criteria:
   - Topic #2: report on 1-2 recent research publications
   - Date: published within last five years
   - Author: no more than 10 texts from any single author

Evaluation: Operational Domain <— Corpus

- Reddit's *r/science* does not include all possible popular science sources
- Scraping procedure does not get a fully random sample
- Research report articles only represents one strand of popular science writing

Appendix B: List of intransitive verbs included in verb pattern analyses

Intransitive verbs included in analysis of SNA corpus (48)
occur, work, happen, increase, end, exist, live, move, change, spread, do, drop, persist, rise, survive, decline, die, disappear, fall, grow, pass, start, stay, travel, emerge, form, improve, lie, result, sit, vary, arise, continue, differ, dissolve, double, extend, orbit, remain, shrink, stop, accumulate, begin, break, decrease, develop, escape, evaporate
Intransitive verbs included in analysis of RA corpus (40)
increase, occur, vary, decrease, differ, exist, emerge, remain, fall, arise, grow, lie, change, move, drop, continue, begin, decline, participate, persist, extend, respond, identify, form, do, start, accumulate, hold, improve, develop, evolve, score, rise, die, overlap, scale, last, live, end, operate

Appendix C: List of multi-word verbs included in verb pattern analyses

Phrasal verbs (150)
set off, keep on, run out, make out, shut up, turn off, bring about, step back, lay down, bring down, stand out, come along, play out, break out, go around, walk out, get through, hold back, write down, move back, fill out, sit back, rule out, move up, pick out, take down, get on, give back, hand over, sum up, move out, come off, pass on, take in, set down, sort out, follow up, come through, settle down, come around, fill in, give out, give in, go along, break off, put off, come about, close down, put in, set about, set up, turn out, get out, come in, take on, give up, make up, end up, get back, look up, figure out, sit down, get up, take out, come on, go down, show up, take off, work out, stand up, come down, go ahead, go up, look back, wake up, carry out, take over, hold up, pull out, turn around, take up, look down, put up, bring back, bring up, look out, bring in, open up, check out, move on, put out, look around, catch up, go in, break down, get off, keep up, put down, reach out, go off, cut off, turn back, pull up, set out, clean up, shut down, turn over, slow down, wind up, turn up, line up, take back, lay out, go over, hang up, go through, hold on, pay off, hold out, break up, bring out, pull back, hang on, build up, throw out, hang out, put on, get down, come over, move in, start out, call out, sit up, turn down, back up, put back, send out, get in, blow up, carry on, go on, find out, come out, go out, point out, grow up, pick up, come back, come up, go back
Prepositional verbs (119)
think about, worry about, know about, talk about, do about, say about, hear about, look after, known as, serve as, regard as, seen as, considered as, defined as, look at, stare at, glance at, aim at, arrive at, get at, laugh at, work at, allow for, account for, apply for, work for, provide for, make for, used for, stand for, call for, look for, go for, wait for, pay for, play for, call for, ask for, differ from, come from, suffer from, make from, take from, hear from, used in, believe in, seen in, occur in, result in, included in, engage in, succeed in, keep in, place in, fall into, run into, enter into, put into, look into, get into, break into, feel like, sound like, look like, speak of, think of, make of, hear of, composed of, consist of, conceive of, know of, work on, go on, put on, concentrate on, spend on, rely on, depend on, base on, get over, go through, get through, turn to, apply to, send to, give to, talk to, write to, speak to, listen to, occur to, contribute to, belong to, lead to, happen to, related to, point to, attach to, reduce to, respond to, explain to, refer to, add to, deal with, play with, fill with, cope with, agree with, begin with, start with, work with, stay with, live with, connect with, associate with, meet with, compare with, involve with

Appendix D: Interview guide for interview study

1. How long have you been writing science journalism for? What do you typically write about?
2. Who is your intended audience? What level of background knowledge do you assume they have?
3. How do you choose what to include and omit when writing from a research article?
4. In my study, research articles used more of the past tense and science news articles relied more on the present tense. Does that seem accurate to you? If so, why might this be?

[PPT SLIDES #1-2]

5. In my study, research articles used a lot of passive voice verbs compared to science news articles. Does that seem accurate to you? If so, why might this be?

[PPT SLIDES #3-4]

6. In my study, science news articles used a lot of modal verbs, especially the word *can*. What might be useful about these words for writing science news?

[PPT SLIDES #5-6]

7. In my study, science news articles were found to often provide topical background information by referencing a general group of people. Why do you think these phrases are useful for science news articles, and how important is providing background information in writing such an article?

[PPT SLIDE #7]

8. Related to the previous question, many of the verbs used in these phrases expressed what scientists *think* or *believe*. Why do you think these verbs are useful when writing science news articles?

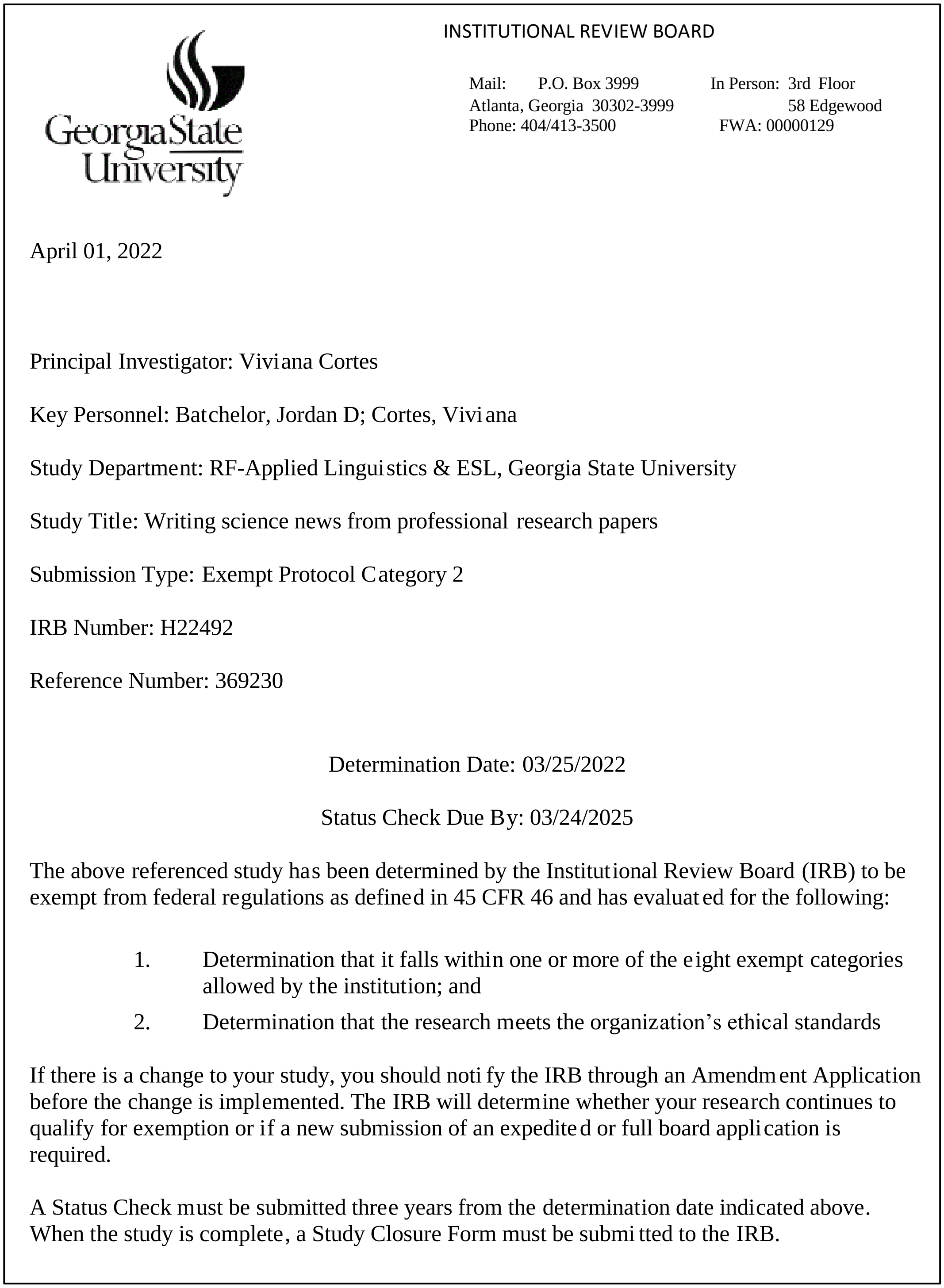
[PPT SLIDE #8]

9. In my study, science news articles used a lot of sentences with a subject, verb, and complement clause. Often, the subject was a person and the verb described what they said or thought, with the complement clause containing what was said or thought. Why do you think sentences like these might be useful for science news writers, and less so for RA writers?

[PPT SLIDE #9-10]

10. Is there anything else you’d like to add or ask me?

Appendix E: Interview study IRB letter



Georgia State University

INSTITUTIONAL REVIEW BOARD

Mail: P.O. Box 3999
Atlanta, Georgia 30302-3999
Phone: 404/413-3500

In Person: 3rd Floor
58 Edgewood
FWA: 00000129

April 01, 2022

Principal Investigator: Viviana Cortes

Key Personnel: Batchelor, Jordan D; Cortes, Viviana

Study Department: RF-Applied Linguistics & ESL, Georgia State University

Study Title: Writing science news from professional research papers

Submission Type: Exempt Protocol Category 2

IRB Number: H22492

Reference Number: 369230

Determination Date: 03/25/2022

Status Check Due By: 03/24/2025

The above referenced study has been determined by the Institutional Review Board (IRB) to be exempt from federal regulations as defined in 45 CFR 46 and has evaluated for the following:

1. Determination that it falls within one or more of the eight exempt categories allowed by the institution; and
2. Determination that the research meets the organization's ethical standards

If there is a change to your study, you should notify the IRB through an Amendment Application before the change is implemented. The IRB will determine whether your research continues to qualify for exemption or if a new submission of an expedited or full board application is required.

A Status Check must be submitted three years from the determination date indicated above. When the study is complete, a Study Closure Form must be submitted to the IRB.

Appendix F: Linear mixed effects model outputs

Appendix F.1: LME model of the relationship between the use of adverbials, register, and topical category

Effect	Estimate	CI	p value	R2
Intercept	42.25	41.29 – 43.20	< 0.001	0.228 R2m 0.393 R2c
Register: SNA	-6.89	-8.96 – 4.81	< 0.001	
Category: Health	2.78	1.64 – 3.92	< 0.001	
Category: Psych	-2.04	-2.70 – -1.39	< 0.001	
Category: Space	-0.44	-0.91 – -0.03	0.067	

Appendix F.2: LME model of relationship between modal verbs, register, and topical category

Effect	Estimate	CI	p value	R2
Intercept	7.22	6.74 - 7.70	< 0.001	0.286 R2m
Register: SNA	5.38	4.67 – 6.09	< 0.001	
Category: Health	-0.67	-1.35 – 0.00	0.05	
Category: Psych	0.74	0.35 – 1.13	< 0.001	
Category: Space	0.31	0.04 – 0.59	0.03	0.308 R2c

Register: SNA* Category: Health	1.79	0.78 – 2.79	< 0.001	
Register: SNA* Category: Psych	-1.60	-218 – -1.02	< 0.001	
Register: SNA* Category: Space	0.43	0.02 – 0.83	0.04	

Appendix G: Raw frequencies of four RC sub-types across three grammatical contexts and two registers

Sub-type	RA		SNA	
	Voice			
	Active	Passive	Active	Passive
Deixis	61	5	348	6
Citation	777	174	343	18
General reference	51	41	150	17
Quotation	16	5	104	3
	Tense			
	Past	Present	Past	Present
Deixis	28	35	164	196
Citation	328	608	118	214
General reference	11	79	42	123
Quotation	6	14	24	80
	Aspect			
	Simple	Non-simple	Simple	Non-simple
Deixis	64	5	362	13
Citation	735	214	321	35
General reference	64	27	118	47
Quotation	18	3	89	17