ASTON UNIVERSITY

DOCTORAL THESIS

Corpus-based Study of the Rhetorical Organization and Lexical Realization of Scientific Research Abstracts

John Blake

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

March 2021

ASTON UNIVERSITY

Abstract

School of Languages and Social Sciences

Doctor of Philosophy

Corpus-based Study of the Rhetorical Organization and Lexical Realization of Scientific Research Abstracts

by John BLAKE

A key difficulty for novice writers when drafting scientific research abstracts is adherence to the discourse community expectation of generic integrity. This is especially the case for writers with English as an Additional Language. In order to meet these generic expectations, it is necessary to understand what the disciplinary norms are in terms of rhetorical move structure and the language used to realize those moves. Therefore, this research aims to describe the rhetorical organization and lexical realization in a corpus of scientific research abstracts.

A balanced corpus of research abstracts (n = 1,000) from top-tier journals in ten scientific disciplines was created. This corpus contained 7,200 sentences, each of which was manually annotated using rhetorical moves as pragmatic units. Tailor-made online annotation tools were created to ensure the quality of the annotations. Specialist informants and double annotators were used to verify annotation accuracy.

Tailor-made scripts were created to identify, count, classify and extract the annotated rhetorical moves. The individual moves within each discipline were counted. The frequencies were compared and contrasted. Patterns for adjacent pairs of rhetorical moves and the full rhetorical move sequences were investigated. The permutations of rhetorical move sequences for each abstract were extracted. These sequence permutations were also compared and contrasted, revealing three dimensions in which differences occur: linearity, cyclicity and variation.

Analysis showed substantial differences in usage among the scientific disciplines. Permutations of move sequences varied in terms of linearity, which was based on the assumption of an expected order of INTRODUCTION, PURPOSE, METHOD, RESULTS followed by DISCUSSION. Some move sequences included the fronting of rhetorical moves, such as placing the RESULT MOVE before the METHOD MOVE. Cyclicity was present in disciplines that were concerned with the development and evaluation of algorithms and artifacts. In these disciplines, adjacent pairs of moves were often repeated, such as METHOD-RESULT, METHOD-RESULT with the first pair of moves describing the development phase and the second describing the evaluation phase. The third dimension was in the variation of permutations found within each discipline. This study found an immense variation in applied scientific and engineering

disciplines, which is in stark contrast to the formulaic abstracts typical in medical research. Slightly under 200 different move sequence permutations were uncovered in the corpus. Analysis was conducted to investigate the similarities and differences in rhetorical organization on the three dimensions of linearity, cyclicity and variation. Based on these results, a *Borromean Rings* framework was devised to map the disciplinary generic conventions of rhetorical organization of research abstracts onto a linguistic landscape with three dimensions. The theoretical implications and practical application of these results are elucidated.

Lexical realization within moves within disciplines was also investigated using keyness and grammatical tenses as proxies for lexis and grammar. Cluster analysis was used to reduce the dimensionality and identify the extent to which keyness and grammatical tense usage are move-specific and/or discipline-specific. The cluster analysis grouped the ten disciplines. The resultant clusters were very similar to the results of analysis using the *Borromean Rings* framework. The main difference being that cluster analysis classified the disciplines slightly more finely in one branch of the dendrogram. Dispersion of key words varied greatly across moves and disciplines.

Both key words and grammatical tenses showed move-specific and discipline-specific collocations and colligations. Disciplinary variation is pervasive, but patterns of collocation and colligations are perceptible. Knowledge of these patterns can help novice writers of scientific research abstracts climb the cline of competence and learn how to draft abstracts that meet the generic expectations of their community of practice.

Acknowledgements

I would like to express my gratitude to my dissertation supervisor, Krzysztof Kredens, for his invaluable guidance and constructive advice that helped bring this project to fruition. I would also like to thank Jack Grieve for his statistical advice and Garry Plappert for the inspiration to use the programming language, R.

I feel especially fortunate to have been employed by universities that not only encouraged this research, but also provided me with the time and resources to concentrate on this. I am particularly grateful to William Holden at Japan Advanced Institute of Science and Technology and to Ian Wilson at the University of Aizu for providing supportive work environments. Special thanks also go to Brian Kurkoski and numerous Ph.D. candidates at the Japan Advanced Institute of Science and Technology whose contributions, comments and suggestions helped bring this research forward in its early days.

Some of the most valuable advice, however, came from discussions with participants and presenters at various academic conferences. Some of these discussions were brief, yet had profound influence on this research. Although it is not possible to list everyone, I would like to name and extend my thanks to those who had the most impact on this study. I sincerely thank (in alphabetical order): Laurence Anthony, Emily Bender, Veejay Bhatia, Christine Feak, John Flowerdew, Lynne Flowerdew, Stephan Th. Gries, Michael Halliday, Michael Handford, Ken Hyland, Cameron Smart, John Swales and Elena Volodina.

I would like to add a note of thanks to the specialist informants and double annotators who not only helped improve the annotation accuracy of the dataset, but led me towards a world of logic, algorithms and deductive reasoning.

Although the focus of this research is on applied linguistics, it was necessary to create a number of software programs to identify, extract and analyze data to address the research questions. Over this project, programming moved from procedural programming using C to extract annotations, to Visual Basic for Applications to clean data, to Perl, to R and finally to Python. While getting to grips with these programming languages, I sought help from numerous sources. I owe the most thanks to the developers and users of Stack Overflow, a community-based question-and-answer website for technical difficulties. Although the list of people who gave me advice and helped improve my code is extensive, those who deserve special thanks include my son Xavier Blake, my colleagues Vitaly Klyuev, Maxim Mozgovoy, Evgeny Pyshkin and Julian Villegas, my student Duc Tran Vu and my Python mentor and friend Simon Pavlic.

I would particularly like to extend my gratitude to another friend, Bryan Beaton, who provided innumerable suggestions on how to increase the readability of my manuscript.

Then, there is the most important of all to thank: my family. I would like to thank my father for his support. Without him, this project would never have started.

I am indebted to my wife, Minako, who allowed me to concentrate on this study and supported me wholeheartedly throughout. A final note of thanks goes to Luka and Lena who played and fended for themselves while I wrote this dissertation instead of spending quality time with them.

Contents

List of Figures

List of Tables

List of Abbreviations

BOT	**BOT**any
CARS	**C**reate **A** **R**esearch **S**pace
EAL	**E**nglish as an **A**dditional **L**anguage
EAP	**E**nglish for **A**cademic **P**urposes
EC	**E**volutionary **C**omputing
ELT	**E**nglish **L**anguage **T**eaching
ERP	**E**nglish for **R**esearch **P**urposes
ESP	**E**nglish for **S**pecific **P**urposes
IAA	**I**nter- **A**nnotator **A**greement
IEEE	**I**nstitute of **E**lectrical and **E**lectronics **E**ngineers
IND	**IND**ustrial Electronics
IMRD	**I**ntroduction **M**ethod **R**esults **D**iscussion
IP	**I**mage **P**rocesssing
IPMRD	**I**ntroduction **P**urpose **M**ethod **R**esults **D**iscussion
IT	**I**nformation **T**heory
KDE	**K**nowledge and **D**ata **E**ngineering
KWIC	**K**ey **W**ord **I**n **C**ontext
LING	**LING**uistics
MAT	**MAT**erials science
MDS	**M**ulti- **D**imensional **S**caling
MED	**MED**icine
MWE	**M**ulti- **W**ord **E**xpression
NES	**N**ative **E**nglish **S**peaker
NLP	**N**atural **L**anguage **P**rocessing
NNES	**N**on- **N**ative **E**nglish **S**peaker
SD	**S**tandard **D**eviation
SFL	**S**ystemic **F**unctional **L**inguistics
SOP	**S**tandard **O**perating **P**rocedure
PIMRD	**P**urpose **I**ntroduction **M**ethod **R**esults **D**iscussion
POS	**P**art **O**f **S**peech
WC	**W**ireless **C**omputing

To my wife Minako and to our children Luka and Lena without whom this thesis would have been completed much earlier.

Chapter 1

Introduction

The beginning of knowledge is the discovery of something we do not understand.
- Franklin Patrick Herbert Jr., American science-fiction author

1.1 Chapter preview

The background of this research is described in Section 1.2, including the historical context of research abstracts and an overview of the genre and language features of scientific writing. This leads into a brief introduction to the specific genre of scientific research abstracts, and the linguistic hurdles that novice writers must overcome to draft abstracts that conform to the generic expectations of the research community. These challenges are even more onerous for novice writers who are non-native English speakers (NNES); or, as J. Flowerdew (2008) terms, those with English as an Additional Language (EAL). The relevant extant knowledge on rhetorical organization and lexical realization is summarized. The gaps in the research literature that this study addresses are pinpointed. An overview of the theoretical background that underpins this research is given, and a declaration of the stance of the researcher is also provided.

Section 1.3 describes in more detail the challenges and barriers that novice EAL writers face when drafting scientific research abstracts. The prescriptive-descriptive disjuncture in the pedagogic literature and the complexity of this genre in terms of lexical complexity, lexical density and information density are described. This is followed by a discussion of the practical difficulties experienced by novice writers regarding rhetorical organization (i.e. which rhetorical moves to use and in which order), and the lexical realization of those moves (e.g. which tenses, modalities and aspects prevail in which disciplines).

Section 1.4 provides an outline of the broad research goal and aims of this study. The research aims are elaborated and justified based on a critical review of the pertinent literature in Chapter 2. The need for further research on the rhetorical organization of scientific research abstracts is identified.

The contributions that this corpus-based investigation of scientific research abstracts adds to the literature are then described. These contributions to the research

literature in terms of novelty, importance and substance are highlighted in Section 1.5.

A summary of the introduction is given in Section 1.6. This is followed by the thesis overview in Section 1.7, which briefly introduces the main focus of each of the subsequent chapters.

1.2 Background

1.2.1 Scientific writing and research abstracts

> *Do not write so that you can be understood, but so that you cannot be misunderstood.*
> - Marcus Fabius Quintilianus (ca. 35 – ca. 100), Roman rhetorician

The use of English as the language for science started with the publication of the *Philosophical Transactions of the Royal Society* in 1665. This journal published a variety of types of articles from letters which were customary at that time to "experimental essays", often credited to Robert Boyle (Lareo Martín and Montoya Reyes, 2007). From this point, structured research articles started to evolve into the dense, noun-heavy texts typical in contemporary scientific discourse. A survey of the INTRODUCTION, METHOD, RESULT and DISCUSSION[1] move structure (IMRD) in health sciences over a fifty-year period showed some use of IMRD in the 1940s. By the 1970s eighty percent of all articles adopted IMRD, and a decade later in the 1980s IMRD was the only format for original research articles (Sollaci and Pereira, 2004). Research abstracts, however, have a much shorter history as they did not appear until the 1960s. The medical disciplines were the forerunners in introducing research abstracts (Swales and Feak, 2009, p.1). Medical disciplines later were also the first to adopt structured research abstracts in which specific components of the abstracts are prescribed. The use of structured abstracts is claimed to improve the peer review process (Hartley, 2014), digital retrieval (Haynes et al., 1990; Hayward et al., 1993) and quality of information (Sharma and Harrison, 2006).

Scientific research abstracts may appear to be a very narrow genre or part-genre (Swales and Feak, 2009) of writing but there is considerable diversity. Abstracts tend to be evaluated by reviewers primarily on originality, importance and substance. The originality could be to address a hitherto undiscovered niche or the development of a novel method or algorithm. The importance could be related to the issue under study or the results themselves. A simple concrete example of importance is a medical breakthrough that could save many lives. Substance refers to the difficulty and rigour of the research.

The taxonomy of research abstracts is complex. Research abstracts can be categorized by publication type, namely: dissertation, conference and journal abstracts.

[1] Small capital letters are used for the names of rhetorical moves

Abstracts can also be classified according to the subject area (e.g. medical or engineering) and content coverage (e.g. informational or indicative). In addition, there are also promissory, structured and graphical abstracts. Combinations of these categories result in a multitude of possible permutations of abstract types, such as promissory linguistic conference abstracts and informative structured medical journal article abstracts. This study concentrates on one of the most prestigious genres: research abstracts of scientific articles published in top-tier journals.

Scientific articles incorporate language features of both research writing and scientific writing. Scientific writing has been described as one of the most important discourse types (Halliday and Martin, 1993; Holtz, 2009). The linguistic features of scientific writing have been the focus of many studies. In addition, there are numerous pedagogic books dedicated to providing advice to novice writers on how to master scientific writing. Some of the distinctive features of both scientific writing and scientific research abstracts include: specificity (Parkhurst, 1990), linguistic complexity (Halliday and Martin, 1993), linguistic density (Holtz, 2009), abstraction and nominalization (Halliday and Martin, 1993; Martin, 1993; Matthiessen and Halliday, 2009), long noun phrases (Biber and Gray, 2013; Gopen and Swan, 1990; Vande Koppel, 1994; Biber, Conrad, and Rippen, 1998) and enhancing the appearance of objectivity (Holes, 1995, p.260) through the avoidance of personal pronouns (Hatim and I. Mason, 1997) and passive voice.

1.2.2 Scientific research abstracts

The editorial guidelines of journals frequently stipulate that research articles should be "prefaced by homotopic abstracts" (Swales, 1990, p.178). Research abstracts have multiple functions. One of the primary functionalities is to act as a gateway to the full research article (Hartley and Benjamin, 1998). Potential readers of research articles frequently narrow down the number of research papers to read using key word searches, titles (Anthony, 2001; Haggan, 2004; Hartley, 2007; Jamali and Nikzad, 2011; Sagi and Yechiam, 2008; Soler, 2011; Y. Wang and Bai, 2007) and abstracts (Amnuai, 2019; Anderson and Maclean, 1997; Can, E.Karabacak, and Qin, 2016; Esfandiari, 2014; Holmes, 1997; Holtz, 2011; Jiang and Hyland, 2017; Lorés, 2004; Pho, 2008; Salager-Meyer, 1992; dos Santos, 1996; Swales and Feak, 2009; Stotesbury, 2003; Suntara and Usaha, 2013; Tseng, 2011; Tu and S. Wang, 2013). Research abstracts, therefore, act as filter to help researchers cope with the ever-increasing amount of published research (Ventola, 1994, p.333). From the viewpoint of potential readers, the abstract should function as an objective summary (Swales, 1990, p.179) that provides the most important details of the content of the research article. However, the viewpoint of the writer may differ. This is because writers of research articles may aim to entice as many potential readers as possible to download, read and cite the full version. Research careers may be contingent on various citation indexes that aim to measure the impact of published research. Scientific research articles can, therefore, be considered as a type of promotional genre (Hyland, 2012b; Samar et al., 2014).

As complex meanings are compressed into as few words as possible, scientific research abstracts tend to be more linguistically and information dense than their associated research articles. The combination of a large number of subject-specific technical terminology and the overall high lexical density makes scientific research abstracts a particularly difficult genre to understand. In comparison, research abstracts in the humanities are easier to understand and to a certain degree non-subject specialists can extract the general and specific details of research abstracts in disciplines they are not familiar with. In some scientific disciplines, lay readers may be able to understand most of the words and make an educated guess of the gist of the research, but it is unlikely that lay readers would be able to describe specific details without detailed subject-specific knowledge.

This challenge is exacerbated for researchers with English as an additional language who need to read and write research abstracts in English. Readers who know the content area, specialist terminology and are familiar with the genre of scientific research abstracts, should be to understand the content of the abstracts while those lacking knowledge in one of these three areas may find that comprehension evades them.

There are a number of difficulties facing authors of scientific research abstracts. Two difficulties are conforming to discourse community expectations (particularly those of the reviewers and journal editors) and developing sufficient lexical and grammatical competence. Swales (1990) asserts that mastering the process of abstract writing may be considered as one of the *rites de passage* for entry into the scientific discourse community.

Drafting research abstracts is a daunting task for many junior researchers, both native speakers and non-native speakers alike, but particularly so for non-native English speakers (NNES) who have to not only master their own discipline but also come to grips with English (Englander, 2013; Hanauer and Englander, 2013; Hanauer, Sheridan, and Englander, 2019). Scientific language is challenging to understand at both the conceptual and linguistic levels. Those with little scientific background in the particular discipline of research are unlikely to understand the intended meaning of the author despite understanding the grammar of the sentences. Conversely, scientists whose first language is not English may understand the concepts, but have difficulty accessing the intended meaning because of the language barrier. With the assistance of tools such as translation software and grammar checkers, researchers may be able to construct an English version of their research abstract. Yet, without sufficient exposure to research abstracts written in English in their specific discipline, the chance of adhering to generic expectations is minimal. A major linguistic barrier to drafting a scientific research abstract is the adherence to the generic integrity of the target publication in terms of rhetorical organization and lexical realization. Although the consequences for violating generic integrity cannot be predicted accurately, novice writers may reasonably expect that their work will not be accepted for publication (Devitt, 2004).

Specifically, novice writers need to know (1) what to write, i.e. which content of their research should be included, (2) in which order should the content be included, and (3) which words and phrases can be used to describe their content in an appropriate manner.

Research abstracts have been the focus of a number of research studies. The research literature on research abstracts that is pertinent to this study can be divided into three domains. First, studies that investigate rhetorical organization, such as Nwogu (1990), Salager-Meyer (1990), Ventola (1994), dos Santos (1996), Samraj (2002), Hyland (2004), and Swales and Feak (2009). Second, studies that investigate disciplinary variation, such as for voice (Melander, Swales, and Fredickson 1997; Yakhontova 1999; Hyland 2004; Samraj 2005; Pho 2008). Third, studies that investigate lexical realizations (Suntara and Usaha, 2013) either within moves or throughout abstracts, such as hedging (Rounds, 1982), tense (Amnuai, 2019; Esfandiari, 2014; Graetz, 1982; Salager-Meyer, 1992; Swales and Feak, 2009; Tseng, 2011; Tu and S. Wang, 2013), and *that* clauses (Hyland and Tse, 2005).

However, despite this apparent firm foundation, the majority of the studies mentioned were conducted in social sciences with a large number of the studies clustering on linguistic disciplines (e.g. Suntara and Usaha 2013) and medical disciplines (e.g. Salager-Meyer 1990; Salager-Meyer 1992; Salager-Meyer 1994; Salager-Meyer, Ariza, and Zambrano 2003).

Some of the studies are corpus-driven, corpus-led or corpus-based. There are, however, three main caveats to this body of research.

First, many of the early studies, particularly ones conducted prior to the 1990*s* adopted an armchair analysis approach. In this approach, the linguistic researcher draws upon their memory of a genre or a particular text or group of texts within the genre, and then uses that introspective analysis as the basis of their study or their interpretation of results from a very small collection of research abstracts. Fillmore (1992) argued cogently for the place for both armchair linguists and corpus linguists, but also proposed a midway position, a computer-aided armchair linguist. At that time, the perceived drawback of armchair linguistics was the lack of empirical evidence to support any claims while corpus linguistics was thought not to be able to discover anything interesting, given its focus on counting the frequency of linguistic features.

Second, research abstracts from more accessible disciplines, particular linguistics, were often chosen. It is unsurprising that linguists chose to investigate linguistics abstracts as the content and format is familiar to them. Fields which attract substantial research funding, such as medical disciplines (National Science Foundation, 2018) are also frequently studied. No doubt the necessity to provide front-line medical professionals with access to the latest research results in the most effective and efficient manner is a strong driver for this. Data mining and natural language processing researchers also focus on medical disciplines. Given that human lives may be saved by enabling medical doctors to extract specific information from medical articles,

these disciplines are likely to remain well funded and well researched.

Third, research on disciplinary variation has frequently shown that even closely-related disciplines differ substantially in terms of language usage (Lorés, 2004; Samraj, 2002). However, numerous factors ameliorate the generalizability of findings from popular disciplines to lesser investigated hard science disciplines, such as information theory and wireless communication. Some of the factors that impinge on the transferability of research findings include: (1) the size of the corpus, (2) the aims of the research, (3) the rigour of the annotation and (4) the lack of reported results on lexical realization within rhetorical moves.

Researchers have argued over how many rhetorical moves occur in abstracts, the necessity or otherwise of particular moves and the degree of variation among disciplines. The comparisons to date have been descriptive along the lines of "abstracts in discipline X tend to include move Z more frequently than in discipline Y". No-one to date has suggested a framework onto which the abstracts of all scientific disciplines could be mapped. Such a framework could allow instances of each abstract to be plotted, and then using statistical analysis, clusters of abstracts could be identified.

Although researchers have focused on rhetorical organization for many years, the research efforts have shed little light on the permutations (i.e. the particular order and sequences of combinations) of rhetorical moves. Almost all research has focused on which moves and sub-moves (e.g. steps or strategies) occur in which disciplines rather than which combinations or particular permutations of moves occur.

Corpus linguists who have annotated specialist corpora tend to complete the annotation of rhetorical moves in one of three ways:

1. annotate the abstracts themselves,
2. cajole their graduate students to annotate, or
3. recruit subject specialists.

Each of which has the potential to introduce significant bias. High values of inter-annotator agreement may provide some ammunition in defence of the method chosen, but these high values rule out unreliability rather than establishing validity.

The first way enables the researcher to maintain quality (or at least consistency) and obviates the need for payment. Self-annotation may explain why so many researchers selected their own discipline of linguistics as the object of study. The second way is popular, particularly for researchers holding professorial positions, but when annotating abstracts in disciplines outside their own discipline, questions may be raised on the accuracy of the annotations. Recruiting subject specialists should result in high quality annotations as they understand the content, with the proviso that they can understand and adhere to the annotation protocol. This option, however, is the most resource intensive in terms of time and finance. For researchers with strong inter-departmental connections within universities and those with higher status within their institution, securing co-operation may be easier and less costly.

When reporting the annotation procedure, almost all published articles have either claimed high agreement or stated simple percentages as proof of inter-annotator agreement, and hence of (perceived) reliability and (assumed) validity. In order to convince readers of the rigour of their research, more sophisticated statistical analysis and more details on the annotation protocol are needed (Blake, 2018).

To date, there is scant research on the disciplinary variation of collocation and colligation within rhetorical moves in abstracts in hard science areas, such as various disciplines in the engineering domain, specifically those disciplines that may be loosely grouped into the multidisciplinary field of information science. The impact of this field is far-reaching. For example, machine learning, neural networks and artificial intelligence have already made substantial inroads into linguistics. To illustrate this, over a third (88/257) of the accepted long papers included "neural" referring to "neural networks" in their title in the 2018 proceedings of the Conference of the Association of Computational Linguistics (Gurevych and Miyao, 2018).

1.2.3 Theoretical background

When reading a text, it is usually possible for readers to classify the text according to its genre. When reading an article, most readers would be able to work out whether the article is from a journal or a newspaper. When reading a letter, most readers can deduce whether the letter is personal, business, official or junk mail. Classifying texts more finely into narrow generic categories is more challenging, but with sufficient exposure to the genres or sub-genres, most readers are able to do so. Genre exists, but there is no agreement among researchers on exactly what genre is. Children who have read many novels and short stories are, in general, more likely to be able to write stories that hold the attention of the reader. However, children with little exposure to stories and novels are unlikely to be able to do so. This holds true for other genres, too. Imagine trying to write a legal contract in a foreign language without ever seeing a legal contract. The task is insurmountable. However, with some familiarity of legal contracts in one language, it is possible to attempt to draft a legal contract in another language. Yet, there is a high likelihood that the contract would not be fit for purpose.

To understand a genre takes extensive exposure or intensive study. Native English speakers tend to understand genres written in English through extensive exposure, but non-native speakers living in countries that do not use English as a *lingua franca* are unlikely to get sufficient exposure without intensive study. Genre analysis can be used to create a framework or recommended phraseologies that help novice writers, particularly those with English as an Additional Language, to draft texts. Genre analysis deals with purpose, audience and message of the language used. Different discourse communities share different values, cultures and ways of expressing meaning. Texts drafted by members of the same community of practice tend to show generic integrity. Core members of the community may opt to not confirm to generic expectations, but novice writers tend to conform to gain acceptance.

Move analysis is the most commonly used method of genre analysis by teachers and developers of courses that teach English for Specific Purposes. A rhetorical move, simply put, is a small chunk of text. The chunk is a discrete unit that fulfils a particular function. Dividing large stretches of texts into bite-size chunks helps learners understand how texts are constructed and provides novice writers with building blocks with which their own texts can be created. Each genre can be analyzed for moves and the sequencing of the moves investigated. Some genres, such as greeting cards, are relatively formulaic while other genres, such as novels, are not. The first step in move analysis is to understand which moves are utilized in texts. The next step is to uncover any patterns in usage. Patterns can include the co-occurrence of particular moves or the sequencing of sets of moves. With knowledge of the moves and their patterns of usage help learners to create an outline or first draft of their text.

A corpus approach was adopted in this study to analyze the rhetorical moves used in published research abstracts. This approach is empirical and evidence-based rather than intuitive and speculative. A balanced representative corpus was created, annotated and analyzed. Annotating a corpus helps add value to a text by focusing on language features that may not be discovered simply through simple rule-based parsing. This allows the discovery of less obvious patterns of linguistic features and interactions between those features, such as the colligation between rhetorical moves and grammatical tenses.

This study adopted a cyclical approach (Wallis, 2007) that combines top-down and bottom-up approaches to corpus analysis in line with suggestions from researchers advocating investigating texts from different perspectives (Bhatia, 1983). The annotation process started with a top-down approach (Biber, Connor, and T. A. Upton, 2007) by focusing on the functional or communicative purpose. The initial annotation scheme was informed by the corpus, and using a bottom-up approach the annotation scheme was adjusted to more accurately reflect the purpose of the research. The application of top-down approaches, such as move analysis, is highly labour-intensive and rarely undertaken for a large corpus of texts. The reward for overcoming the difficulty of applying an analytical framework, such as a detailed annotation protocol, to a large corpus is the ability to extract detailed analyses. Frequency-based analyses can reveal typical structures, such as move sequences, which would not have been possible without such annotation. These sequences could be used by teachers of research writing to provide frames, models or examples that novice writers could experiment with when creating their own scientific research abstracts.

1.3 Problem statement

The first step in solving a problem is to recognize that it does exist.
- Zig Ziglar, American author

One key problem facing novice writers of an unfamiliar genre is maintaining generic integrity (Bhatia, 1993). These writers need to master not only their own

discipline, but also the genre of writing scientific research abstracts that conform to the expectations of their peers in their scientific discipline. This hurdle becomes more substantial when operating in a second or additional language, and that difficulty is exacerbated by the degree of linguistic distance (Chiswick and P. W. Miller, 2005) between the mother tongue or primary language of the writer and English. For example, writers with first languages of French, German or Spanish face far fewer language difficulties than those with first languages of Chinese, Korean or Japanese.

The topic of writing research abstracts is the focus of few books but is regularly addressed in books geared towards helping researchers draft articles. Many such books provide generic advice geared towards no specific scientific discipline. One early example of such book is "An outline of Scientific writing" (J. Yang, 1995, p.53) which states:

> ...an abstract should answer the questions, *why*, *how* and *what*. *Why* did you study it? *How* did you study it? *What* did you find and *what* does it mean?
> *Why* can be omitted if the objective is clear in the title. *How* should be elaborated on only if it is a paper on methodology; otherwise, it should be very brief or even omitted if well-known. *What* should selectively include only the important findings and conclusions.

The majority of the books provide little insight into the variety of types of abstracts, and with the notable exception of Swales and Feak (2009) fail to provide advice that is based on corpus findings. The distilled version of the generic advice is: write abstracts in a linear format sequencing moves starting from INTRODUCTION MOVE and finishing with the DISCUSSION MOVE; yet when skimming through engineering abstracts, it becomes obvious that there is some discrepancy between the rhetorical organization used in practice and that presented in textbooks.

Lexical and grammatical complexity add further difficulties for EAL novice writers. When writing research abstracts, it is necessary to select not only expressions that convey the precise meaning, but also to select words appropriate for the context in terms of meaning, form and appropriacy. When writing scientific abstracts, the most precise concise terminology is frequently selected. The high number of technical terms whose meanings are opaque to non-specialists, combined with a copious amount of assumed knowledge tends to result in texts are clear to the target audience of specialists. Hyland (2006, p.22) notes that this complexity excludes lay readers and creates a linguistic entry barrier to the academic community. Swales (1997, p.374) coined the term English as a *Tyrannosaurus rex* to describe when English is used as a linguistic tool of power like a "carnivore gobbling up the other denizens of the academic linguistic grazing grounds". The creation of dense scientific texts is achieved through nominalization (Biber and Gray, 2013) which also is one of the causes that make such texts impenetrable to non-specialists.The lack of redundancy in language used adds yet another hurdle for EAL writers and readers who are at a linguistic

disadvantage. Plavén-Sigray et al. (2017) found the readability of research abstracts steadily decreased in all scientific disciplines from the late nineteenth century until 2015 due to increases in general scientific jargon, severely limiting the accessibility of research findings to lay readers. Decades earlier researchers, such as Snow (1963), were already expressing concern at the mutual incomprehensibility between scientific disciplines (Fuller, 2005, p.39).

Rhetorical organization has been the subject for a number of corpus investigations. However, most of these studies focus on a limited number of scientific disciplines with linguistics and medicine being the most commonly researched disciplines. While disciplinary variation is not a new concept, few linguistic researchers have ventured into the more technical disciplines to understand how rhetorical moves are actually used. The entry barrier is, perhaps, rather high, given the necessity to gain sufficient knowledge of the subject discipline to understand the gist of an abstract, and more detailed knowledge to understand the specific details. This hurdle no doubt explains the paucity of research describing these highly technical disciplines. With no first-hand experience of such texts, a naive assumption would be that these disciplines follow the rhetorical organization patterns that are described in the pedagogic literature. Only by investigating a corpus can this assumption be confirmed or refuted. Although the investigation of rhetorical moves *per se* is not novel, no study has specifically focused on the sequencing patterns of these moves.

Once the rhetorical moves are identified, it is possible to work out which functional exponents are used in different disciplines and different moves. The extent to which functional exponents differ among and between moves and disciplines has not been studied in depth for a broad spectrum of scientific disciplines. With a clearer understanding of how rhetorical moves are realized, more accurate advice could be shared with novice writers. Writers with EAL would benefit greatly from access to such information.

1.4 Research aims

> *No great discovery was ever made without a bold guess.*
> - Isaac Newton, English mathematician

This research aims to provide a comprehensive descriptive account of the rhetorical organization in scientific research abstracts published in top-tier journals. This study covers a broad range of scientific disciplines with a particular emphasis on the under-researched disciplines that may be grouped under the umbrella term of information science.

The linguistic goals are to describe and explain the rhetorical organization of scientific research abstracts and to identify the degree to which lexical realization of rhetorical moves differs among moves within and among different disciplines.

A general theoretical goal is to identify a framework linking rhetorical moves and subject disciplines. This framework should show which disciplines share similar

patterns of rhetorical organization and should be able to accommodate any scientific discipline, not just those investigated in this study. This framework is not intended to be prescriptive, but to serve as a model onto which data can be mapped to show the linguistic landscape of this important genre of scientific writing.

A practical goal is to help novice writers by providing them with a schemata on which they can base the initial drafts of their research abstracts. This helps establish a base on which writers can build their own abstract by choosing to follow established conventions or deviate from them when they feel justified. When drafting, it is necessary to make rhetorical and linguistic choices at the level of rhetorical moves. Writers would benefit greatly from knowing which rhetorical moves to include in their abstracts, and in which order this information is usually presented. In addition, knowing which grammatical tenses and which phraseologies predominate in different moves and disciplines may reduce the time that it takes for novice writers to select the most appropriate expression or grammatical tense to convey their message. Collocation and colligation are extremely difficult for writers with insufficient exposure to their target genre, and so providing guidelines to help writers select appropriate language features would aid such writers.

1.5 Contribution to literature

> *Theories are like a stairway; by climbing, science widens its horizon more and more, because theories embody and necessarily include proportionately more facts as they advance.*
> - Claude Bernard, French physiologist

This research contributes to the theory on rhetorical organization and lexical realization within rhetorical moves. This theory can be applied to help developers and teachers of scientific research writing provide advice that aligns more closely to the descriptive reality than the prescriptive advice proffered in textbooks.

This result contributes to the body of literature on research abstracts in three ways. First, this is the first large-scale corpus of research abstracts comprising a range of scientific disciplines that was manually annotated for move structure. Statistical analysis of this corpus showed the plethora of permutations of move sequences shedding light onto a hitherto under-researched aspect of rhetorical moves. Second, on further analysis, patterns in the move sequencing in different disciplines were identified, namely the presence or absence of cyclicity, linearity and variation. These three aspects have to date not been focused on in the published literature on research abstracts. Third, using n-grams to investigate the lexical realization with moves, it was possible to show the effect that move and discipline has on language selection. To the best of my knowledge, no one has explored the rhetorical organization of scientific research abstracts to the same depth and breadth as reported in this study.

This evidence-based description primarily focuses on how rhetorical moves are organized. The practical ramifications for the results are wide ranging. First and

foremost, the discovery of a plethora of permutations of rhetorical structures is particularly pertinent and flies in the face of generic advice that is proffered to novice writers.

As mentioned above, the concepts of cyclicity, linearity and variation can be harnessed by teachers of writing to help writers draft scientific research abstracts that meet the generic expectations of their respective discourse communities of practice.

The framework developed in this study can be utilized to show how disciplinary variation affects the rhetorical move structure. Both teachers of research writing and novice writers of research abstracts benefit from this knowledge. This is in line with Hyland (2002, p.113), who notes pedagogic materials ought to be based on "analyses of representative samples of the target discourse".

The results of this corpus-based, genre-analytic study can help teachers of research writing develop research-supported materials incorporating suggestions based on the corpus evidence of both the rhetorical organization of moves and the lexical realization within those moves. This research is the first to describe in detail the disciplinary variation within a broad range of scientific research abstracts, including less-studied disciplines, such as image processing, information theory and evolutionary computation.

Novice writers with English as an additional language can also directly benefit from the application of this research in two ways. Both of these ways strive to achieve what Lee and Swales (2006, p.72) termed "technology-enhanced rhetorical consciousness-raising". First, the key outcomes of the research could be distilled into small multimodal website designed to help writers draft scientific research abstracts. This website could provide the actionable advice based on the corpus findings. The advice can include the type and common sequences of rhetorical moves that predominate in particular disciplines. A case in point is the cycling of METHOD and RESULT MOVES in industrial electronic and wireless communication. Second, novice writers could access an online tool that visualizes rhetorical moves in context. Users would select a corpus of research abstracts in their discipline and explore the corpus, revealing and hiding the move annotations to better understand which moves are used in which order. This hands-on data-driven learning could give them the opportunity to see at a glance the permutations of moves, since each move is shown in a different colour. This "noticing" according to Schmidt (2012) may or may not be conscious but without noticing, language acquisition may not occur.

1.5.1 Importance

This section shows that there is a need for an in-depth exploration of scientific research abstracts in order to help NNES researchers disseminate their research more effectively and contribute to science using English as a *lingua franca*.

The transition away from Latin to English as the learned language of scholarly publication in the sciences in Europe occurred from the 17th century to the end of the 19th century . Since then, English has become the *de facto* language of articles

published in scientific journals (Simionescu and Simion, 2004). English is likely to remain *in situ*, given that "a third of the world's population, some two billion persons, now use English" (Britannica, 2013). English is therefore highly likely to continue to be the dominant language in science and technological research in the foreseeable future (Graddol, 1997; Montgomery, 2013)).

From 2000 to 2018, the number of worldwide users of the internet grew by over 1000% to over 4.2 million giving a penetration rate of over 55% (Miniwatts Marketing Group, 2018). In 2017, there was an estimated 7.8 million researchers in the fields of science and engineering according to the UNESCO science report (UNESCO, 2018, p.14). This figure does not include the substantially larger number of undergraduate and graduate students who also read but are less likely to write research abstracts. Research abstracts is the genre that is the most commonly read by scientific researchers, and so clearly deserve to be the focus of research.

The importance of scientific research abstracts is shown by exploring their functionality, the personal motivations for writing them and the potential professional benefits to authors (Hayer et al., 2013, p.352). To quote Hoffmann (2010, p.312), "virtually all of a scientist's work will be judged first (and often last) based on an abstract". Research abstracts have become one of the most prominent genres in scientific writing (Swales and Feak, 2009). For scientists working in academia, not only are abstracts a summary of the research undertaken, but they are also an essential element in the pursuit of publishing in prestigious journals and presenting at international conferences, both of which enhance the curriculum vitae of the researcher. This, in turn, increases opportunities for tenure, promotion and securing funding. In short, researchers are likely to be instrumentally motivated to draft abstracts to increase not only the readership of their research, but also to support their "cycle of credibility" (Latour and Woolgar, 1986, pp.200–1), improving both their professional and financial standing. No doubt this instrumental motivation explains why Latour and Woolgar (1979) (also cited in Hyland, 2004, p.3) argue that scientists in research laboratories spend more time and energy on producing research papers than on making discoveries. The proportion of time dedicated to writing could be higher for researchers who use English as an additional language.

There has been a phenomenal increase in the number of scientific publications (Kirkpatrick, 2009), the availability of research abstracts online, and the volume of research output, particularly in the field of information science. The scientific output is growing substantially with output from third world countries increasing dramatically (Blickenstaff and Moravcsik, 1982). The growth rate for scientific publication has been at least 4.7% per year, meaning that publication volume doubles every 15 years (Okulicz-Kozaryn, 2013). This phenomenon has been described as an "information explosion" (Swales and Feak, 2009, p.1). It is also notable that the number of articles authored by non-native English speakers is increasing (Ferguson, 2007, p.43; Montgomery, 2013, pp.89–96).

The move from bound journals to electronic publication means that research

abstracts are now easily found using search engines. Google Scholar is a particularly powerful search engine enabling users to find abstracts through any net-enabled device. There has been an exponential increase in the number of abstracts for conference proceedings and scientific journals that are available online. All open-access journals and most pay-for-access journals publish research abstracts online, making them freely available to not only the scientific community, but anyone with access to the internet. Abstracts are, therefore, of even more importance to attract readers. Additionally, since web crawlers, such as Googlebot, access the full-text of abstracts, word choice and syntax can radically affect the search engine keyword ranking results whereas previously the keywords were simply selected by the author.

1.5.2 Novelty

The under-researched disciplines are more technical, and include some of the most rapidly growing and important disciplines, such as Information Theory, which provides the underpinning theory behind innovations in encryption, wireless communications and many other technical disciplines. Despite their importance, these disciplines have not attracted much attention from linguists or corpus linguists. The most likely reason is the intrinsic difficulty of understanding the meaning of the research abstracts. Without the underlying background in the subject area, the meaning of the research abstracts is likely to evade most readers. This is due to the combination of highly technical terminology in conjunction with advanced scientific concepts. The lack of a conceptual framework of the subject, the lack of familiarity with the technical vocabulary combined with the lack of awareness that the lay words, such as *get*[2] and *call*[3], may have alternative technical meanings, no doubt contributes to this.

L. Flowerdew (1998) noted that the trend in corpus-based analyses was moving towards examining corpora at not only the lexical and grammatical levels, but also at function and pragmatic levels. Annotation of discourse-pragmatic functions, such as rhetorical move structure, remains rather rare, and tends to be in the realm of computational linguistics and natural language processing rather than corpus or applied linguistics (Gries and Berez, 2017). This relative rarity helps demarcate the niche of this research.

Another novelty of this research is the focus on the sequencing of rhetorical moves. As will be shown in the literature review, although rhetorical moves have gained much attention, no research study was uncovered that focused on the sequences of moves in research abstracts. Most studies focused on the sets of moves and their status as optional or obligatory. This new perspective enables rhetorical moves to be considered not simply as a discoursal units, but as patterns permeating texts.

[2] *Get* is a method of requesting a resource via a browser.
[3] *Call* means to invoke a routine in a programming language.

1.5.3 Substance

This research project involved in-depth investigation of a corpus of 1000 scientific research abstracts. In total 7200 sentences were manually annotated with the rhetorical moves and sub-moves. This in itself was no small undertaking given the complexity of the content and technical nature of the language in the texts.

The sequences of the rhetorical moves were identified, and the sequence patterns were counted and classified. This research involved a substantial time cost in terms of both human annotation hours and in the development of the research methodology and tailor-made software to achieve this task. Further analysis was undertaken to uncover the proportion of potential permutations of rhetorical move sequences that were actualized in the corpus. To better understand the sequencing of rhetorical moves, analysis was also undertaken for adjacent pairs of moves to investigate how moves combine together in particular sequences. Once the rhetorical moves were annotated, the lexical realization within the moves could be investigated. A tailor-made script was developed to automatically identify twelve verb forms commonly taught in EFL textbooks. These verb forms will be referred to as *grammatical tenses*[4]. The grammatical tenses were analyzed by rhetorical move and by discipline to ascertain any move-specific and discipline-specific patterns of tense usage. Key words within rhetorical moves were also investigated to discover whether move-specific and discipline-specific patterns could be perceived.

This study was, therefore, non-trivial. A substantial amount of time had to be dedicated to understanding the relevant terminology and basic concepts of each scientific discipline in order to complete the annotation. To identify the twelve grammatical tenses multiple prototypes of tense detection and identification programs were developed, which involved significant linguistic analysis. To perform the analyses, tailor-made software had to be developed. The development process was substantial and involved the creation of numerous functions.

1.6 Chapter summary

The genre of research abstracts is relatively new with a history of less than a century. Although there are many different types of research abstract, this study focuses on research abstracts published in top-tier journals. Research abstracts for scientific articles axiomatically display typical features of scientific writing, but the necessity to convey information concisely creates a genre of writing that is linguistically dense and complex. This combination of lexical density and complexity makes scientific research abstracts a challenging genre to read and write. Lay readers for some of the more technical disciplines are unlikely to be able to understand either the underlying concepts or the technical terminology, creating an insurmountable barrier to comprehension.

[4]Grammatical tenses include both modal and tensed verb forms.

English is the *de facto* language for scientific research, and given the necessity for scientists to publish in English, English research abstracts are a high-stakes genre that can make or break careers.

Linguists have investigated the rhetorical organization in a number of scientific disciplines. However, the investigations have been primarily in the easier-to-read disciplines, such as linguistics, medicine and social sciences. The applicability of these results to the more technical scientific research abstracts is unknown.

The theoretical underpinning of this research draws on genre analysis, particularly rhetorical move analysis pioneered by John Swales, and developed by Veejay Bhatia and Ken Hyland. Most of the research on rhetorical moves has focused on the presence or absence of particular moves or sub-moves, resulting in claims regarding the obligatory or optional nature of moves. However, this research aims to describe the actual move sequences discovered in published abstracts.

Problems that novice scientific writers face when drafting research abstracts are deciding what to include, in what sequence should they be arranged and what functional exponents to use. Writers need to be aware of the generic expectations of their discourse community so that they can maintain generic integrity. Unsurprisingly, this task is more challenging for writers with little exposure to the genre and with less proficiency in English.

Many books and websites provide advice to writers, but the extent of the applicability of the generic advice to the technical scientific disciplines is unknown. It is also unknown whether the advice proffered for rhetorical organization in the pedagogic literature is accurate.

This study aims to describe the rhetorical organization within a multidisciplinary corpus of scientific research abstracts. Specifically, focusing on the sequencing patterns of the rhetorical moves. With a corpus of scientific research abstracts annotated for rhetorical moves, the lexical realization can be investigated. The current lack of description and discussion of the effect of scientific discipline and rhetorical move on lexical choice is under-researched.

1.7 Thesis overview

Chapter 2 reviews the literature on genre analysis, rhetorical move analysis and scientific research abstracts. The literature review then focuses on the rhetorical organization and lexical realization within scientific research abstracts. The chapter culminates with the research questions (Section 2.8) designed to address the research gaps discovered in the review.

As a corpus linguistic approach was chosen, Chapter 3 provides an overview of corpus linguistics and a succinct argument for the adoption of this approach. The benefits and drawbacks of corpus approaches are itemized and evaluated. The theory and practice of selecting, annotating and analysing corpora are discussed.

Chapter 4 describes the methodology used to find the answers to the research questions. A three-phase approach was adopted: corpus phase, annotation phase and analysis phase. The methods used in each of the three phases of the study are described in detail.

The result and discussion chapters are integrated. The chapters are grouped topically by the research questions and sub-divided into sections following the sub-research questions. The penultimate section of each of these chapters itemizes and describes the theoretical implications and practical applications while the final section provides a concise conclusion for each topic.

Chapter 5 is the first result-and-discussion chapter, which presents the results relating to rhetorical organization. The results are organised, described and discussed following the sequence of the related six sub-research questions detailed in Section 2.8.

Chapter 6 is the second result-and-discussion chapter, which describes the findings that relate to lexical realization. The results for each of the remaining four sub-research questions are given, explained and discussed.

The conclusion (Chapter 7) brings together the various threads and arguments that were developed in the combined result-discussion chapters. Some of the key takeaways from this project include the discovery of a prescriptive-descriptive disjuncture between the pedagogic literature and the corpus of scientific research abstracts; a framework that maps disciplinary abstracts based on their rhetorical organization; and patterns of permutations of rhetorical moves among different disciplines despite the dominance of discipline-specific terminology.

Chapter 2

Literature review

> *The greatest part of a writer's time is spent in reading, in order to write: a man will turn over half a library to make one book.*
> - Samuel Johnson. The Life of Samuel Johnson, LL.D. Volume 2.

2.1 Chapter preview

This chapter synthesizes, summarizes and directly relates the research literature to the focus of this dissertation, namely investigating the rhetorical organization and lexical realization in scientific research abstracts across a range of scientific disciplines.

This chapter begins by introducing and defining genre in Section 2.2. The three main schools of thought on genre theory are compared and contrasted. The pervasiveness and complexity of genre, and intricate links between genre and the community of practice are highlighted to contextualize the production of texts by and for the scientific community. This section shows how academic disciplines may be viewed as academic tribes, and the need for novice writers to maintain generic integrity to join the community of practice is detailed. The concept of generic integrity is explained in depth. Genre analysis from a linguistic standpoint is discussed, and the concept of a communicative or rhetorical move is introduced.

The functional units, rhetorical moves, are dealt with in more detail in Section 2.3. Rhetorical moves, move analysis and move determination are discussed in turn. This section notes the widespread use of move analysis in the English for Specific Purposes (ESP) community. Moves may be determined automatically using pattern matching, with probabilistic pattern matching algorithms that utilize machine learning generating the best automated results. However, as of the date of writing, manual identification of rhetorical moves is substantially more accurate. Given the complex nature of the form-function relationship, empirical and statistical support for moves *per se* in the literature is lacking. Move analysis, nevertheless, continues to exert its dominance over other genre analysis methods, particularly in ESP.

The specific genre of scientific research abstracts and its importance to the scientific community is explored in Section 2.4. The concomitant rise in internet access, internet users and online journals alongside the default status of English as the language of science has led to a proliferation in scientific research abstracts written in English

even for research articles published in other languages. English is shown to be the *de facto lingua franca* for the scientific community. One of the key criteria for success in academia is to secure publication in top-tier journals. This means that research articles and their associated abstracts are high-stakes genres that can make or break careers.

The extant literature on rhetorical organization of scientific research abstracts is reviewed in Section 2.5. This section shows that there is currently no research that provides a comprehensive picture of scientific research abstracts in terms of the rhetorical organization of moves. Studies have discovered which moves are frequently realized, and which of those moves are obligatory. However, there is little research that provides insights into the permutations or sequences of moves. This gap in the literature provides a niche worthy of further study.

The lexical realization within moves in scientific research abstracts is described in Section 2.6. This section reveals the concomitant paucity of research on lexical realization within moves in scientific research abstracts. The main focus is on the literature related to lexis and grammar using keyness and grammatical tense as access points to investigate patterns of usage. Although a number of studies have investigated key words in abstracts, only a handful of small-scale studies have looked at lexical realization within moves. No large-scale multidisciplinary study has yet to deal with lexical realization within moves. A few small-scale studies have mentioned the usage of tense and aspect in research abstracts. No large-scale cross disciplinary studies that report on tense and aspect used systematically within and across rhetorical moves could be found in the extant literature.

The penultimate section (Section 2.7) summarizes the literature review drawing on the main aspects raised in each individual subsection, focussing on the gaps in the body of research on the rhetorical organization and lexical realization of scientific research abstracts.

Section 2.8 gives the two main research questions that arise out of gaps that were identified in the research literature. The first research question stems from the section on rhetorical organization 2.5 while the second research question focuses on discovering more about the lexical realization 2.6, specifically the key words and grammatical tenses used within and across moves and disciplines.

2.2 Genre analysis

> *I think genre rules should be porous, if not nonexistent.*
> - Kazuo Ishiguro, British novelist

2.2.1 Overview

This section discusses the literature on the interrelated concepts of the genre, community of practice and generic integrity. This is followed by a brief overview of genre

analysis, starting with the contributions of Swales and Bhatia who were the pioneers in the field of genre analysis in the field of English for Specific Purposes (ESP).

An introduction is given in Subsection 2.2.2 starting with its broadest form and then narrowing down to genre applied to writing in public, private and hidden genres. Some of the multiple meanings ascribed to genre by scholars are discussed and a definition appropriate for this study is selected. The main schools of genre theory are briefly compared and contrasted in 2.2.3, contextualizing the approach selected in this study.

Subsection 2.2.4 focuses on communities of practice. In this research, the community centers around particular discourses, and so the term discourse community is also used to refer to the same community, i.e. the scientific community that produces and consumes genres of scientific research. The inextricable links between community of practice and genre affect meaning potential in the texts and add an addition level of complexity to any analysis.

The term generic integrity, coined by Bhatia, is defined and discussed in Subsection 2.2.5. The need for members of a discourse community and in particular new entrants to a discourse community to adhere to the expected generic integrity is considered.

Subsection 2.2.6 focuses on genre analysis for pedagogical purposes, such as the analyses undertaken by teachers and materials developers of ESP courses. The concepts of *move, sub-move, step* and *strategy* are defined. A six-step framework proposed by Bhatia to analyze texts is also described. Justification is given for the selection of move analysis, the dominant method of genre analysis in the field of ESP.

A concise summary is provided in Subsection 2.2.7, which asserts that move analysis, considered in Section 2.3, is the dominant method for genre analysis, particularly in the fields of teaching writing and ESP.

2.2.2 Genre

Genre: Opera to occluded texts

Grammar is an organization system that when applied to discourse describes the system in terms of entities (e.g. parts of speech) and their relations (e.g. syntax). Grammar may also be applied more formally to describe, for example: architectural styles, e.g. Palladian villas (Stiny and Mitchell, 1978) and used to reconstruct archaeological sites in virtual reality (Rodrigues et al., 2008).

Genre like grammar may be applied to non-linguistic entities. Genre may be considered as an organization system for verbal and non-verbal discourse (Frow, 2014, p.1). Films, art, television and websites may all be categorised into specific genres. Television viewers can easily categorize programmes into news, documentary, comedy and so forth. More finely-grained classification, however, takes more in-depth knowledge of both the content and the taxonomy of the genre. Genre spans from the broadest classification schemes differentiating, for example, opera from

ballet and novels from speeches. Each genre domain may be further divided into more specific categories, and so novels may be identified as science fiction, fantasy fiction and so forth. Each of these may be further divided into subgenres. Time travel, space opera and post-apocalyptic are some of the subgenres of science fiction.

Genre in this study is limited to texts, specifically scientific research abstracts. However, a broader understanding on the genres of written communication helps to contextualize this study. One school that has investigated genre in depth for pedagogic purposes is the Sydney school.

Hyland (2007, p.153) described the Sydney school of genre as having the most "clearly articulated approach to genre both theoretically and pedagogically". The Sydney school adopts a "stratal model of language in social context"(Rose, 2013b, p.209) and focuses on the social semiotic and functional linguistic aspects of communication. A comparison of the Sydney school and other genre schools can be found in Subsection 2.2.3. A factor impacting on genre is context, which in the systemic functional linguistics can be described through register or metafunction (Halliday, 1985; Halliday, 1994). The genre-based literacy methodology that stemmed from SFL approach in the Sydney school is based on interacting with and guiding students through collaboration on a joint project or shared experience using a three-phase approach: deconstruction, joint construction and individual construction (Rose, 2013b, p.222).

Genres may be public. private or hidden. Public genres are those that can be easily accessed. For written texts, public genres include billboard posters, newspapers and research articles published in scholarly journals. Two areas which are heavily researched are professional and research genres.

Veejay Bhatia dedicated his career to researching professional genres. He extended the work of Swales (1990) to professional contexts (Bhatia, 1993), particularly legal contexts. His work on generic integrity and the assertion that "[a]ll disciplinary and professional genres have integrity of their own, which is often identified by reference to a combination of textual, discursive and contextual features" (Bhatia, 2013, p.241).

One particularly insightful visual he published in Bhatia (ibid.) is reproduced in Figure 2.1. This shows how text is affected by genre which is, in turn, affected by professional practice, all of which are situated in the professional culture.

A particularly notable professional genre is that of research publications. A scientist employed by a university who writes a research article could be considered to operate in three genres: scientific writing, academic writing and research writing. However, distinguishing between these types of writing in a scientific research abstract is not possible. Although research articles have long been the focus of genre studies, the lack of an operational definition or criteria for demarcating such articles was noticed and the tacit understanding of the generic qualities of a research article was codified by Enk and Power (2017).

Private genres are those that are not usually freely shared. Given the trend to digital texts, however, the boundary between private and public genres is rather

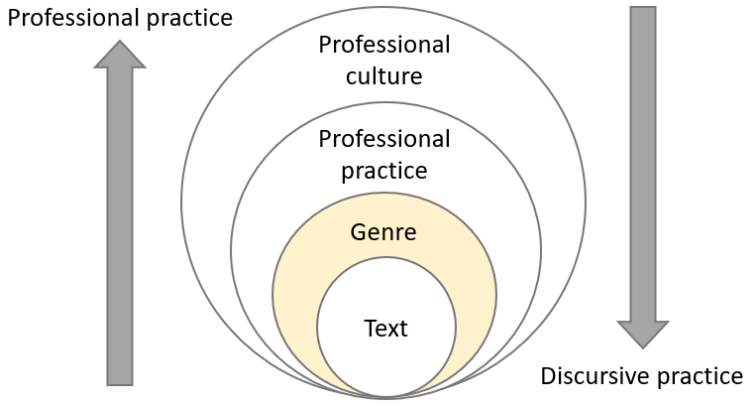


FIGURE 2.1: Perspectives on professional genres
Source: Based on Figure 17.2 in Bhatia (2013, p.248)

fuzzy. Hidden genres include the occluded genres that are rarely shared. Some occluded genres in academia include grant applications (Connor and Mauranen, 1999; L. Flowerdew, 2016), submission letters (Swales, 1996) and rejection letters (Farley, 2017; Jablin and Krone, 1984; Yakhontova, 2019). Forensic linguists also deal with occluded genres, such as suicide notes (Galasinski, 2019) and threats (Bojsen-Møller et al., 2020).

Genre: Definitions

> *Form ossifies into genre through repetition.*
> - Rumaan Alam, American novelist

Swales (1990, p.33) in his seminal book *Genre Analysis* asserts that "genre is a fuzzy concept". Twenty-five years later, genre remains amorphous in nature. His focus is on the use of language for a particular purpose, which could be considered the broadest definition of genre. Swales (ibid., p.58) describes genre as follows:

> A genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by the expert members of the parent discourse community and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style.

Paltridge (2006, p.84) provides a succinct easy-to-understand description of genre: "ways in which people get things done through their use of spoken or written discourse". The functional aspect (purpose or goal) is also present in definitions of genre in the field of systemic linguistics in which genre can be viewed as a "staged, goal-oriented social process" (Hasan, 1996; Martin, 1992; Martin and Rose, 2007, p.8).

Rose (2013a) views genre as occupying the final place on a cline from reading, to text, to text type, to genre. Simply put, this means that individuals read a text to extract

meanings. Those meanings are not context-independent, but are co-constructed not only between the reader and the text, but among texts of a similar type, which as a group may be assigned to a particular genre. The concepts of genre and text type appear closely related. Although there is no consensus on the exact meaning of these terms, these terms can be disambiguated. Text has been defined by as a stretch of spoken or written language, which is semantically and pragmatically coherent in its context (Carter and M. McCarthy, 2006).

O'Donnell (2013) insightfully suggests that genre may be defined in two ways: internal (sharing a common social purpose) or external (sharing common linguistic styles and structure). Martin and Rose (2007, p.7) link genre and meaning as follows:

> Since patterns of meaning are relatively consistent for each genre, we can learn to predict how each situation is likely to unfold, and learn how to interact in it.

And, in doing so, focus on the internal criterion. Martin and Rose (2012) note a disjunct between theoretical genres and texts by stating:

> The apparent gap between ideal genres and real-world texts is in fact an inherent feature of social semiosis as a communication system. Genres flexibly adapt themselves to co-textual, inter-modal and contextual environments as needs arise; and some adaptations, if recurring often enough, give rise to new genres as a culture evolves. (pp.15–16)

In line with Grieve et al. (2010), in this study, the terms genre and text type will be differentiated according to the internal or external focus. Genre is used to refer to the variety of language that is defined by the external situation in which the language is used. This language is characterized by conventionalized linguistic features, and text type is used when referring to a type of language defined exclusively by linguistic properties (Biber, 1995; Biber, 1989).

2.2.3 Schools of genre theory

Genre theory is based on the underlying view that "texts can be classifiable and have understandable predictable forms, structures and purposes" (Knapp, 1997, p.113). Hyon (1996) first proposed the existence of three main schools of thought on genre theory (see also: A.M. Johns, 2002; Hyland, 2003), namely: ESP, systemic functional linguistic (SFL) and New Rhetoric. Yanchun (2007) (also cited in Suntara, 2013) provides a tabular summary of the similarities and differences among the three genre schools. An abridged and amended version is given in 2.1. J. Flowerdew (2002, p.91) proposed a simpler dichotomy of two camps within these genre schools: linguistic (ESP and SFL) and the non-linguistic (New Rhetoric). Both the ESP and SFL genre schools take a "text-first" approach while the New Rhetoric genre schools takes a "context-first" approach (ibid.; Paltridge, 2006, p.98). This dichotomy, however, was

contested by Coe (2002, p.197) who explained that focus of the New Rhetoric school is on the "functional relationship between text type and rhetorical situation". A.M Johns et al. (2006) note that the differing definitions, different starting points (i.e. text or context) and different theoretical and pedagogic emphases make categorization complex.

TABLE 2.1: Similarities and differences among the three genre schools

Aspect	ESP	New Rhetoric	SFL
Defining criteria	Communicative purpose	Recurrent social actions	Goal-orientated purposeful activity
Social use of content	Discourse community	Community ownership	Context of culture
Perspectives on text	Genre shapes the schematic structure of the discourse and constrains the choice of context and style.	Genre knowledge includes both form and content and a sense of what is appropriate to a particular purpose at a particular point of time.	Genre is concerned with systems of social processes: the ways in which field, mode, and tenor are phrased into each other; these variables converge on texture.
Medium of analysis	Texts	Users and context	Text
Unit of analysis	Move and step	Chronotope	Text and strata
Research methods	Analysis of text	Case studies, interview, observation, protocols	Analysis of text

Based on Yanchun (2007, p.18)

Each of the genre schools uses rather different terminology to describe similar concepts. Their theoretical bases differ with their respective purposes. The English for Specific Purpose (ESP) school is primarily concerned with pedagogy and how best to help non-native learners of English understand and produce language that is appropriate for their professional purpose. The starting point of the researcher in that school is the text or texts. The New Rhetoric genre school adopts a more social stance and focuses on the language users *per se* and the manner of their responses in different rhetorical situations. This is the reason their method of analysis starts with the users and uses chronotope as the unit of measure. Chronotope is used in Bakhtin (1981) to show how time and space are represented and realized in language. The systemic functional linguistic approach to genre analysis is also social, but focuses on paradigmatic description using a stratified social-semiotic system. This system was devised by Halliday (1994) and aims to describe texts written in any language by analysing the paradigms, which combine together to create a model of the network of language choices. Genre or text analysis usually begins with the metafunctions of ideational meaning (or phenomena), interpersonal relations and textual. The textual metafunction is how the ideational and interpersonal relations are bound together. Text and context are common themes to all the genre schools.

2.2.4 Community of practice

> *A biologist, if he wishes to know how many toes a cat has, does not "frame the hypothesis that the number of feline digital extremities is* 4*, or* 5*, or* 6*," he simply looks at a cat and counts. A social scientist prefers the more long-winded expression every time, because it gives an entirely spurious impression of scientificness to what he is doing.*
> - Anthony Standen

Gross (1996, p.13) asserts that "innovation is the *raison d'être* of the scientific paper" but continues and notes that the paper must "invoke the authority of past results" and be "embedded in a network of authority relationships" Gross (ibid., p.27). One means of invocation is through citation, the use of which varies by discipline. In one study, Hyland (2004) found that citations were twice as frequent in hard sciences as social sciences. Scientific writing and argumentation style show variation among different disciplines and different cultures (Clyne, 1991; Galtung, 1979; Halliday and Martin, 1993; Markkanen and Schröder, 1992).

Lave and Wenger (1991, p.98) coined the term "communities of practice" to define those involved in the production of texts over time. However, Barton (2007) proposes that discourse community is a looser coupling of individuals involved in the production and/or reception or a text (pp.75–76). Faigley (1986, p.535) claims that writing is understood from the perspective of society rather than that of the individual, and so for an individual to create a message that is understood by the readership, it is necessary to be familiar with the societal perspective. Faigley's society could be termed community of practice or discourse community. Bhatia (2014, p.213) notes that all forms of discourse and especially those used within institutions are "socially constructed and negotiated", which again situates text production as a collaboration. Swales (1998) uses the term "textography of communities" to describe those participating in creation of discourse within specific academic disciplines. Hyland (2004, p.2) claims that "sanctioned social behaviours, epistemic beliefs, and institutional structures" are revealed in the academic writing of different disciplines (see also Bazerman, 1988; Berkenkotter and Huckin, 1995; Myers, 1990). Hyland (2004, p.3) continues "as discourse is socially constitutive rather than simply socially shaped; writing is not just another aspect of what goes on in disciplines, it is seen as producing them". Relating these ideas to scientific research abstracts, Pho (2008, p.231) notes the importance of the acquisition of the skill of abstract writing for novice writers to "enter the discourse community of their discipline". "[N]ewcomers [are] socialized into the practices of members" (Hyland, 2006, p.20) and with this socialization are able to move along the continuum from novice to expert, which in Lave and Wenger (1991) would describe as moving from the periphery to the core of the community of practice.

Different academic disciplines may be viewed as academic tribes (Bartholomae, 1986; Becher and Trowler, 2001; Bourdieu, Passeron, and Saint Martin, 1996; Hyland,

2006, p.13; Swales, 1988). Each tribe has its own symbols, stories and rituals, which need to be mastered to signal membership of the tribe (Bourdieu, Passeron, and Saint Martin, 1996). For example, applied linguists in the corpus camp, tend to use terms such as corpora, node and n-gram, while those in the psycholinguistic camp are more likely to utilize terms, such as mental lexicon and lexical decision. These lores need to be mastered before full acceptance into the tribe is achieved, for example by publication in a peer-reviewed journal. Numerous researchers in various fields have described how members in a group adopt the behaviours and values of a group. Bartholomae (1986, p.4) describes how students have to "invent the university for the occasion"; Swales (1990) and Becher and Trowler (2001) describe the different discourse communities as "tribes" since each tribe has its own distinct differences. C. Miller (1994) asserts that a culture may be characterized by its genre set. Anthony (1998, p.82) quotes his specialist informant who describes research article writing as "preaching to the cannibals", showing the power that the gatekeepers of the community of practice (i.e. the editor and reviewers) exert over a community (i.e. those seeking to get published). Bartholomae (1986, p.4) notes that the acquisition of academic literacy is primarily concerned with joining the discourse community. "[Students have] to learn to speak our language, to speak as we do, to try on the peculiar ways of knowing, selecting, evaluating, reporting, concluding, and arguing that defines the discourse of our community."

One way researchers can weave their way towards the core of the community of practice is by adhering to the practice of citing earlier works, reusing phrases and phraseology that other members use while avoiding accusations of plagiarism.

Citations are a form of intertextuality in which texts refer implicitly or explicitly to knowledge or words in other texts. Fairclough (1992, p.117) classifies intertextuality into two categories: manifest intertextuality (ways of referring, e.g. quotation, paraphrase, summary) and interdiscursivity (use of generic conventions, associated with institutional and social meanings and power). In a similar vein, Lillis and Curry (2010) focus on textual construction dialogic in academic publishing with an emphasis on the contributions of reviewers and editors who despite significant contributions to some texts are not listed as authors. In a number of cases the published research abstract may differ significantly from the initially-submitted abstract at the start of the review process. According to Myers (1991), the construction of scientific knowledge is not an individual activity undertaken by the author alone, but a joint activity in which the author, colleagues of the author and gatekeepers to the publication (specifically reviewers and editors) and (perhaps most importantly) the providers of research grant funding all contribute to the final draft. The resultant publishable version could be argued to balance the author's contribution with the diverse aims of community. Whether this enhances or diminishes the quality of the research is a moot point, since all peer-reviewed publications of funded research are subjected to the similar processes.

There are a number of difficulties facing authors of scientific research abstracts.

Notable difficulties include lexical and grammatical competence, the ability to write plagiarism free, and conforming to discourse community expectations, particularly those of the reviewers and editors. Echoing Bartholomae (1986), Swales (1990, p.187) asserts that mastering the writing process may be considered as the *rite de passage* of entry into the scientific discourse community. Bereiter and Scardamalia (1987) propose different models for novice writers and skilled writers, namely a knowledge-telling model and a knowledge-transforming model. Skilled writers actively rework their thoughts, idea and the text while novice writers tend to focus on content generation. Flower (1979) notes that immature writers tend to create "writer-based" text while mature writers create "reader-based" text, i.e. orientated to the perspective of the reader. Okamura (2006) interviewed thirteen Japanese research scientists and concluded that a key difference between established writers of research documents was their ability to identify their audience, which is a sign of understanding the community of practice of those who create and read texts. Silva (1993) states that skilled writers compose differently from novices, and use more effective planning and revising strategies. Effective writers anticipate counter-arguments and aim to proactively avoid negative critiques by reducing opposition to their statements. Hyland, Huat, and Handford (2012) describe two key reasons for rejecting the statements, namely failure to meet adequacy conditions (i.e. not corresponding to world view) or acceptability conditions (i.e. not considering the affective expectations).

Aside from barriers to publication such as the lack of access to libraries and the internet (Canagarajah, 1996); novice researchers and, in particular, non-native English speaking (NNES) researchers face formidable challenges to having their first scientific paper published in a top-tier journal. Lying on the periphery of the discourse community, researchers in some non-English-speaking environments appear to be at a distinct disadvantage (Casanave, 1998; J. Flowerdew, 2000; Hyland and Hamp-Lyons, 2002). Cultural expectations regarding the degree of involvement of the writer and reader are thought to vary according to language. Hinds (1987, p.143) asserts that in English the writer is responsible for communicating clearly. However, in Japanese it is the responsibility of the reader to extract meaning accurately (Martín and Burgess, 2006). Clyne (1987) similarly notes that in German the responsibility is on the reader. One realization of this difference in English is, as Hyland (2005b) explains, that writers in English make more use of metadiscourse signals to guide the reader by labelling, previewing and structuring the text.

2.2.5 Generic integrity

> *When I see a bird that walks like a duck and swims like a duck and quacks like a duck, I call that bird a duck*
> - James Whitcomb Riley, American poet (1849–1916)

Texts are not produced in a vacuum and so the social context needs to be considered. Social context is related to a common social purpose. The *raison d'etre* of

scientific communities of practice is to share knowledge. It may be argued that this is an *ostensible* reason given that less altruistic motivations may be at play due to the rewards that publication may offer, e.g. prestige, promotion and tenure. This knowledge sharing aspect of research dissemination can be conducted orally through presentations and discussions, but research articles that are published in journals and conference proceedings are the most valued. Peer-reviewed top-tier publications are the most prestigious. Bazerman (1997, p.19) states that:

> Genres are not just forms. Genres are forms of life, ways of being. They are frames for social action. They are environments for learning. They are locations within which meaning is constructed. Genres shape the thoughts we form and the communications by which we interact. Genres are the familiar places we go to to create intelligible communicative action with each other and the guideposts we use to explore the unfamiliar.

Scientific academic writing is a form of rhetoric designed to construct a particular world view (Bazerman, 1988) through the representation of the process of science as linear (Berkenkotter and Huckin, 1995; Latour and Woolgar, 1986, p.369; Schickore, 2008, p.323). It also worth noting that research articles of non-Anglophone authors may deviate from Anglophone norms in their usage of language with non-standard grammar and non-idiomatic language (Montgomery, 2013, p.97); but they do not deviate from accepted rhetorical organization patterns (ibid., p.99). Bhatia (1994, p.61) initially proposed the term "generic integrity" to describe a text that has a recognizable structure (e.g. moves) and characteristic features (e.g. register and vocabulary).

In Bhatia (2014, pp.143–150) a list of indicators that contribute to generic integrity is provided. The indicators are divided into two aspects: text internal and text external. The text internal aspect comprises contextual, intertextual and textual. The textual component is further divided into lexical, rhetorical-grammatical and discoursal. The text external aspect comprises discursive procedures, disciplinary culture and discursive practices. Discursive procedures is concerned with the contributors, authority, participatory mechanisms and interdiscursivity. Disciplinary culture relates to the generic norms and conventions, professional goals and objectives, and professional and organizational identity. Discursive practices is subdivided into choice of genre and communicative modes.

Texts that do not conform to community expectations are likely to be rejected by the community. In this way communities can assert their values on texts and texts can be created to reflect their values. The term genre knowledge can be used to describe the potential to participate in genres. In the words of Berkenkotter and Huckin (1995, p.ix), genre knowledge is defined as:

> We use the term genre knowledge to refer to an individual's repertoire of situationally appropriate responses to recurrent situations - from immediate encounters to distanced communication through the medium of print, and more recently, the electronic media.

Research abstracts that are written in an informal register are likely to be judged negatively as the generic integrity of register has been violated. Conforming to the unwritten expectations of generic integrity of a particular journal is a challenge for peripheral members of the discourse community who may not have access to a suitable mentor to point out the unwritten tacit knowledge that authors need to acquire. Members at the core of the community of practice, who are already established as *worthy* scholars may feel less need to conform to the expectations and may even challenge the generic conventions.

Short genres with very rigid structures (e.g. research article titles) are easier for newcomers to master than longer genres with less rigid structures (e.g. long research articles and dissertations). For example, in an attempt to maintain generic integrity this dissertation portrays this corpus-based study in a linear manner to guide the reader through the phases in research in an easy-to-understand manner. Should the dissertation have been written to reflect the reality; the early annotations, multiple pilot studies and trials, and false starts would need to be included, but that would detract from the main thrust of the study. Texts that do not comply with expectations are likely to incur the wrath of members of a discourse community. In the same manner, a doctoral dissertation that fails to lead the reader along the expected linear path may place the candidate under considerable strain in the *viva voce*.

Hoey (2001) notes that writers draw upon existing knowledge to frame their contribution to the literature, and thus co-construct the text with an envisaged audience (e.g. the readership of the journal) and more concretely co-construct the text with joint authors, reviewers and editors. However, for those with English as an Additional Language, language instruction and mentoring are more likely to be needed to learn how to frame their contributions and begin their journey to the core of the discourse community. Familiarity with a genre enables writers to draft and readers to extract information more quickly. For example, when skimming and scanning a research article for the conclusion, the conclusion may be stated in the abstract, the introduction and/or the discussion/conclusion depending on the type of research article and the expectations of the readership. Experienced readers are likely to go directly to the appropriate section while novice readers may not. This knowledge of generic integrity can be acquired over time through exposure to the rhetorical organization and associated lexical realizations.

2.2.6 Genre analysis

To help novice writers climb the learning curve faster and understand the unwritten rules that need to be mastered to produce texts that adhere to the expected generic

integrity, it is necessary to become familiar with the format, form and function of the target texts. Genre analysis is a way in which researchers systematically analyze texts for generic features. The focus of this study is on the linguistic side of genre analysis of texts. The texts themselves are such a central focus in the work of teachers and materials developers in ESP that Dudley-Evans (2000) argues the centrality of genre analysis to a theory of ESP. ESP practitioners tend to be eclectic in their selection of methods and activities to enable learners to produce texts.

The so-called fathers of genre analysis in the ESP tradition are Swales and Bhatia, both of whom use the term *move* to refer to functional units. Swales first proposed the term *communicative move*, but in the literature the terms *move*, *rhetorical move* and *communicative move* appear to be used interchangeably. Swales (1990) identified a unit smaller than move, a sub-move, that he named *rhetorical step*, which is usually referred to as *step*. The terms *communicative move* and *rhetorical step* are commonly used in the ESP discourse community. The implication of the term *step* is that the sub-move occurs in a particular sequence. Bhatia (2001) coined the term *strategy* to refer to a sub-move which does not have any connotation of sequence. Lewin, Fine, and Young (2001) used the term *acts* to refer to sub-moves, following Sinclairian terminology (Sinclair and Coulthard, 1975) which no doubt was influenced by the speech act theory of Austin (1962).

Various patterns of rhetorical moves have been identified in stretches of text, such as problem-solution, which Hoey (1983) found in academic, business and social genres and Feak, Reinhart, and Sinsheimer (2000) found in legal texts. This does not mean that solutions cannot precede problems, but in the cases when that happens, it would be contrary to reader expectations, and would therefore be used to create a particular effect or focus. Other common two-move patterns include general-specific (Hoey, 2001) and claim-justification (Henry and Roseberry, 2001).

On the practical side of conducting a genre analysis, Bhatia (1993, pp.22–34) proposed six steps as shown in 2.2 to implement when analyzing a genre. Swales is, however, credited with founding *move analysis*, while his protege Bhatia further developed and refined the theory. One such refinement was the removal of the obligation or implication of sequence from sub-moves, which results in a theory that is more robust to criticism. Rhetorical move analysis is dealt with in more detail in Section 2.3.

2.2.7 Section summary

In summary, genre is a pervasive, complex and "fuzzy" concept (Swales, 1990, p.33). Despite the multitude of definitions and interpretations and lack on consensus on the definition of genre (Bawarshi and Reiff, 2010; Biber and Conrad, 2009, pp.21–23; and Hyland, 2009, p.26), there is consensus that texts can be ascribed to a genre. To quote Bawarshi and Reiff (2010, p.3),

TABLE 2.2: Six steps to understand when analyzing a genre

Step	Action
1	Select a representative text.
2	Harness background knowledge and textual clues to contextualize genre in terms of audience, purpose and message.
3	Compare the text to similar texts to ensure that it is representative of the genre.
4	Study the institutional context to understand its conventions, e.g. site visits, guidelines, interviews.
5	Decide the focus on the study (e.g. moves, lexis, cohesion, etc.) and analyse the text.
6	Use specialist informant to check analysis to confirm findings and insights.

Source:Bhatia (1993, pp.22–34)

> [T]he term genre itself remains fraught with confusion, competing with popular theories of genre as text type and as an artificial system of classification. Part of the confusion has to do with whether genres merely sort and classify the experiences, events, and actions they represent (and are therefore conceived of as labels or containers for meaning), or whether genres reflect, help shape, and even generate what they represent in culturally defined ways (and therefore play a critical role in meaning-making).

The language used in a genre reflects the shared goals, audience and type of message. For example, scientific research abstracts are, in general, easily identifiable as research abstracts by scientific researchers. Given the pervasive nature of and importance of genre, it is worthy of further investigation. The interaction between the discourse community of scientific researchers and the texts they produce is also complex. The genre of the texts evolves as innovations occur in the communities, yet the genre, or more specifically the interpretation of the genre by those involved in the peer review process, may also limit such innovations. Core members of the discourse community can innovate and break away from generic expectations, but novice writers are unlikely to be accepted should they fail to maintain generic integrity. This necessity for conformity is reflected in tribal lore of academia. Thus, conforming to generic integrity more easily enables researchers to get published yet by conforming the meaning potential is constrained and innovative approaches to documenting research are held back.

Given the lack of agreement on what genre is, axiomatically genre analysis is fraught with even less consensus. The lack of consensus, nevertheless, has no correlation to the importance of genre analysis. Genre analysis can be used to provide frameworks, templates, lists and exemplars that can help both novice and experienced writers better understand the texts in focus. When used for pedagogical purposes, particularly ESP, move analysis is the dominant method of genre analysis especially in the field of teaching writing as a second language. It should be noted that the SFL school is particularly influential in teaching genre-based writing in the mainstream Australian school system. The concept of move is developed further in the following

section. This is followed by a discussion of move analysis and its application to writing, in particular, to help writers of English as an Additional Language.

2.3 Move analysis

Too much knowledge and analysis can be paralysis.
- Alejandro GonzÍez Inarritu, Mexican film director

2.3.1 Overview

This section provides a contextualized introduction to rhetorical moves, approaches to move identification and modes of move identification in relation to analysis of scientific research abstracts. Move analysis is necessarily closely related to moves, since without analysis moves may not be identified, and conversely without moves, move analysis is not possible. The Swalesian concept of moves is described and discussed in Subsection 2.3.2 using scientific abstracts as examples. Rhetorical moves that are firmly established in the literature are described. The pertinent literature on the complexity of the form-function dilemma is analyzed, and the concept of prototypical rhetorical moves is exemplified.

The following subsection (Subsection 2.3.3) examines bottom-up and top-down approaches to move identification in detail. Subsection 2.3.4 considers the benefits and drawbacks of human annotation and automated annotation. Freely-available software for move identification, namely Mover and MAZEA, is evaluated. Machine and human move annotation are compared, contrasted and evaluated. Subsection 2.3.5 summarizes the issues raised regarding moves, their annotation and analysis.

2.3.2 Rhetorical moves

In this subsection rhetorical moves will be described in greater detail, beginning with providing a definition and an overview of the labels assigned to moves in the research literature. The complexity of rhetorical moves will be explored at both semantic and syntactical levels. Prototypical and ambiguous rhetorical moves are shown to provide greater insight into the non-trivial task of identifying moves in context.

Definition of rhetorical move

A move is, in the words of Swales (2004, p.228), "a discoursal or rhetorical unit that performs a coherent communicative function in a written or spoken discourse". Example 1 announces the result of proving something while Example 2 provides detail about the method used to obtain results.

(1) In this paper, we prove that there does not exist a binary self-dual doubly even code with an automorphism of order 9.
[IT 013]

(2) Numerous experiments are conducted in this paper to demonstrate the performance of our approach.
[IP 1]

The concept of rhetorical moves has been used in teaching English as second language for over 30 years. The strong influence of Swales (1990) solidified their usage within ESP circles. While the model of moves and steps is established and accepted in many fields of research, some criticism still persists on its subjective nature and the lack of empirical evidence in support of the model, but far outweighing this academic criticism is the practical applicability of the concept of move.

Rhetorical moves make an ideal pedagogic tool, since they enable large texts to be divided into functional discrete teachable units. This concept of dividing the whole into parts is advocated during teacher training when trainee teachers are advised to switch from holistic to serialistic approaches to help learners cope with large stretches of text. A typical classroom technique is to show the whole text, divide the text into parts for examination and then revisit the whole text. The concept of moves simplifies the operationalization of this technique since in the serialistic phase, teachers can help learners understand the vocabulary and grammar used in the rhetorical moves.

Teachers and materials developers can then focus students on discrete moves and the specific functional exponents and related vocabulary that may be used to realize such moves. Move analysis has been adopted extensively in both the business writing and the ESP communities. As such, moves are widely used by front-line teachers. Although the term *move* is not always present in the course books and materials, the division of texts into moves is commonplace in both business writing and ESP writing courses.

Established rhetorical moves

The research literature has shown that some rhetorical moves are frequently used. Their rank in the move and sub-move hierarchy varies slightly depending on the researchers and their research purpose. Moves which frequently occur in the tag sets used for annotation of research abstracts (e.g. Hyland, 2004; Pho, 2008; Salager-Meyer, 1990; Swales, 1990) include:

1. BACKGROUND
2. DISCUSSION
3. GAP
4. INTRODUCTION
5. METHOD
6. PURPOSE

7. RESULTS

The names of rhetorical moves are capitalized to enable the name of the rhetorical move to stand out and not be confused with other potential meanings, such as the name of a section of a research article.

These moves can be organized into a hierarchical structure by rankshifting some moves into sub-moves. In this way, the standard IMRD organization can be used with three moves being subordinated into the INTRODUCTION MOVE. This could result in the following tag set:

1. INTRODUCTION
   (a) BACKGROUND
   (b) GAP
   (c) PURPOSE
2. METHOD
3. RESULTS
4. DISCUSSION

This resultant tag set is somewhat of a hybrid between the traditional IMRD model and the Create-a-research-space (CARS) model. The CARS model is a three-move model developed by Swales (1990) that was created to describe the rhetorical organization of scholarly research studies, specifically introductions. The three moves within CARS are:

1. ESTABLISHING CENTRALITY
2. ESTABLISHING A NICHE
3. OCCUPYING A NICHE

A more in-depth discussion of the CARS model can be found in Subsection 2.5.3.

Complexity of rhetorical moves

> *The grammatical system has ... a functional input and a structural output.*
> - Michael Halliday

The relationship between linguistic form and communicative function is complex. It is known that the cardinality of a one-to-one relationship has been ruled out (Lassen, 2006), and so if there is correlation the relationship is many-to-one or one-to-many. This relationship, however, may vary with respect to the size of the ontological unit.

One dilemma in annotation of functions is the rift between form and function. The meaning of a word can be semantically divided into denotation (core meaning) and

connotation (associative meaning). Connotation is what makes it easy to differentiate nuances between terms, such as thin, slim and skinny. However, despite this apparently simple distinction, there is also a difference between the dictionary meaning of a word and its usage. Although words can be used to mean the same as a dictionary definition of the word, words can also be used to create new meanings by relying on the reader to use context of its use to work out the intended meaning. The differences in meaning and usage are exacerbated when considering larger grammatical units, such as phrases, clauses and sentences. The meaning intrinsic in each word may not combine to reflect the intended meaning of an utterance or sentence. The ability for a word, phrase, clause or sentence to realize different functions gives users of the language an infinite number of possibilities to express themselves. However, decoding the intended meaning or meanings of an expression becomes more problematic. In fact, ambiguity resolution is a topic that is of intense interest to researchers in natural language processing. In short, there is, therefore, a disconnect between the form of the language and the function of the language.

Researchers (Bhatia, 1993; dos Santos, 1996; Samraj, 2005)) have noted that moves are sometimes embedded within sentences. Hyland (2004, p.73) also notes that moves may be merged into one sentence, e.g. Purpose-Method sentences. In a similar vein, Pho (2008) asserted that two different moves may be realized by one sentence. In contrast, in a pilot study of 100 abstracts from five disciplines that are part of multidisciplinary information science, no embedding of moves was found (Blake, 2014).

Move identification and classification is context-dependent and depends on both the rhetorical purpose and the linguistic choices (Connor and Mauranen, 1999). To classify moves accurately both endogeneric knowledge, i.e. that contained in the text itself and exogeneric knowledge, i.e. that not contained in the text may be needed. The exogeneric knowledge comprises world knowledge, which is shared (to a large extent) by humans living in civilisations and discipline-specific knowledge, which is particular to each scientific discipline. As a single ontological unit may consist of one or two moves, it is necessary to create a protocol for assignment of moves. This may be to assign the most dominant move or assign two moves to a single ontological unit.

Prototypical and ambiguous rhetorical moves

In this research rhetorical moves which one would expect all human annotators to accurately identify are classed as prototypical moves while moves which the human annotator may be unable to accurately identify are classed as ambiguous rhetorical moves. This rather simple dichotomy illustrates the problem of second-guessing the intention of the author and also goes someway to explain why machine learning approaches may fail to identify moves accurately.

Let us consider this fictitious overly-simplified four-move abstract.

M1: In this paper we present Algorithm D.
M2: Algorithm B is able to classify Item X to a precision of 0.94.
M3. We use Algorithm C to create a new Algorithm, D.
M4: Algorithm D is able to classify Item X to a precision of 0.96.

Of these four moves, M3 could be considered prototypical as we can reasonably expect all trained annotators to classify this as METHOD MOVE. However, M1 might be coded as PURPOSE MOVE or RESULT MOVE. The syntax of the sentences in M2 and M4 is identical. M2 is, in fact, BACKGROUND SUB-MOVE while M4 is the RESULT MOVE. However, it is not always obvious to non-subject specialists which algorithms already exist and which are the products of the research being described. The lack of exogeneric knowledge inhibits the ability of the annotator to classify the move. Consider the scenario in which researchers developed algorithm B, then used algorithm C to create algorithm D? Then, both M2 and M4 could be argued to be labelled as RESULT MOVES.

2.3.3 Approaches to move identification

When investigating rhetorical organization, a top-down approach begins by identifying purpose to categorize (i.e.create categories and classify) or classify (i.e. classify into predetermined categories) the text into moves (Biber, Connor, and T. A. Upton, 2007). Whereas, a bottom-up analysis focusses on the content and function first (Osborne, 2004), Lieungnapar and Todd (2011, p.2) note that top-down analysis processes predominate in genre analysis. Bottom-up processing, however, can be conducted using linguistic features to ascertain the content and function. Lieungnapar and Todd (ibid., p.2) also reveal in their analysis of one section (the list of contents) of ten academic journals the results from bottom-up and top-down approaches differed. Their sample size of ten texts was too small to make generalizations, but it could be an indication of the difficulty of using linguistic form to predict communicative function. Wallis (2007) describes three types of methodological stances towards the roles of annotation and corpus data, namely:

1. a top-down deductive approach in which knowledge is codified in the annotation scheme,

2. a bottom-up inductive approach in which generalizations are made from the corpus, and

3. a cyclic approach in which knowledge is embedded in both the annotation scheme and the corpus.

In the cyclic approach the observations inform the scheme.

2.3.4 Modes of move identification

Move identification can be carried about by humans who read the texts and label moves. Alternatively, pattern matching approaches, such as rule-based parsing and machine learning, can be harnessed. Rule-based parsing approaches can reach high levesl of accuracy in narrow domains; but, in general, tend to be outperformed by probabilistic approaches using machine learning (Sagae and Lavie, 2003) and so this section will focus on machine learning.

Machine annotation

Automatic machine annotation of moves can be achieved by training a classifier on an accurately annotated training set. The classifier harnesses probabilistic parsing and machine learning to determine which categories to classify the ontological units as being a member of. There are many advantages to this automated annotation, notability the speed, scalability and consistency. The classifier can assign each stipulated ontological unit to the category that the classifier selects in a fraction of a second, and unlike human annotators, classifiers can annotate millions or billions of words in minutes. In its favour, automatic annotation has a higher degree of consistency since given the same data sets and the same parameters, the algorithm will result in the same decision each time (assuming no randomization is set). A drawback to the current state-of-art automatic annotation classifiers is their reliance on endogeneric knowledge, i.e. the knowledge contained within a text. When real-world or discipline-specific knowledge is needed to disambiguate a choice, the automated system is unable to do so.

The focus of NLP researchers is on creating pipelines to process language more efficiently and more effectively. In order to process language, texts and annotations are used. Text categorization and classification involve processing language and assigning texts or segments of text to thematic, topical or functional categories. Machine learning is frequently used to classify texts (see Sebastiani, 2002 for an overview). Machine learning classifiers are often trained on one corpus and tested on a held-out corpus that contains texts that are very similar in terms of genre. The quality of the output is directly related to the quality of the input, and when annotations are inaccurate, the classifier will surely be inaccurate. When annotations are accurate, the classifier has a higher probability of accuracy.

A key problem in move determination is the selection of which linguistic features to use to determine membership of a particular category. In this area, researchers use abductive[1] and inductive reasoning to narrow down the choice of linguistic features. Machine learning, however, harnesses statistical reduction and measures of association such as regression to establish which language features predominate in which classes. This can only be determined to the degree of accuracy of the training

[1] Abductive reasoning attempts to find the simplest and most likely conclusion to explain a set of observations

data. The first set of training data needs to be manually coded, and so any errors in that set may have a cumulative effect on future iterations of annotation. Cotos and Pendar (2016, p.93) note that:

> despite considerable advances in NLP [natural language processing], analyzing the various aspects of natural language is still a challenging problem that remains to be solved in order to meet a wider range of learning needs. One such need is mastering the writing conventions of academic genres.

During the period this research was undertaken, two tools were available to automatically identify moves, namely *Mover 1.0* developed by Anthony and Lashkia (2003) and *Multi-label Argumentative Zoning for English Abstracts (MAZEA)* developed by Dayrell et al. (2012). *Mover 1.0* was designed to annotate the move structure in research introductions automatically using a modified CARS model. Anthony and Lashkia (2003) state that an average accuracy rate of 68% was achieved during their evaluation of annotation accuracy. The second system *MAZEA*, which is no longer available online, was designed to annotate the move structure in research abstracts using six moves: BACKGROUND, GAP, PURPOSE, METHOD, RESULT and CONCLUSION. The developers claim the system is reasonably satisfactory, but noted that it was less accurate on the lexically-dense corpus of physical sciences and engineering abstracts.

However, in exploratory studies by the author neither tool was able to discriminate sufficiently between moves with accuracy rates of less than 50% being commonplace. There were also a number of logistic problems with each tool. These included the difficulty of exporting results from *Mover 1.0* . The *MAZEA system* was only available via an online interface. This necessitated inputting each text individually, waiting while the text was annotated and the copying and pasting the results into a plain text file. As of the time of submission, the online interface for *MAZEA* is no longer available.

Human annotation

Human annotation scores poorly when judged on speed, scalability and consistency. However, the main criterion is none of those factors, but accuracy. Human annotation requires an annotator to read a text, in this research, a scientific abstract, and decide which parts of the abstract realize which moves. Assuming that the categories of moves are already established, then the annotator needs to classify each ontological unit as realizing one or more moves. Human annotation of rhetorical moves is an extremely time-intensive process (Anthony, 1999; Anthony, 2001; Dayrell et al., 2012; Swales, 1990).

The annotator has to second guess the intention of the writer. In many cases, this may be obvious, but at times a lack of exogeneric knowledge, such as world knowledge or discipline-specific knowledge, may result in a frame of reference that is not shared with other annotators.

For some disciplines in which technical terminology may act as a significant barrier to lay readers, it may be necessary for subject specialists to either annotate the texts directly or, at the very least, provide advice on the annotations.

Human versus machine annotation

The lack of consistency for human annotation is due to subjectivity. The intra- and inter-annotator reliability scores are far lower than for machine annotation. This is not to say that machine annotation is objective, though. Subjectivity occurs in both modes. The subjectivity for human annotators occurs in the decision made for each annotation instance while the subjectivity for automatic annotation is created by the software developer prior to annotation as it is built into the training data and the parameters that are encapsulated in the program.

Both automatic and human annotation approaches could be used as the first step with the final classification being decided by the alternative approach. However, it is time-saving to harness automatic move determination first and then use human annotation to increase the precision since manual annotation currently has a higher degree of inter-annotator accuracy.

2.3.5 Section summary

The concept of move is eminently utilizable as a pedagogic tool to divide texts into smaller chunks of text. This division into moves makes it easier for teachers and materials developers of ESP courses to focus on language form and function of these discrete units within the context of the whole text. This method of dividing and analyzing is common in education, and is readily employed by front-line teachers and materials developers.

Moves are not decontextualized, but are firmly situated with the context and co-text of a text. In this study, rhetorical moves within the genre of scientific research abstracts are the object of study. The following section introduces this genre by describing the importance of scientific research abstracts, the status of English as the language of choice, and the taxonomy and types of research abstracts.

2.4 Scientific research abstracts

> *I used to hate writing assignments, but now I enjoy them. I realized that the purpose of writing is to inflate weak ideas, obscure poor reasoning, and inhibit clarity. With a little practice, writing can be an intimidating and impenetrable fog!*
> - Bill Watterson, quote from Calvin and Hobbes

2.4.1 Overview

This section shows the importance of scientific research abstracts from both the perspective of readers and writers. The use of English as the *lingua franca* in the scientific community, the rapid increase in online publishing and the primacy of English language publications all contribute to their importance

Subsection 2.4.2 begins by arguing that the academic prestige of journal articles and the status and hegemony of English as the default language of science together to heighten the importance of research abstracts written in English. The advent of easier access to the internet brought about a sea change in the dissemination of research findings and the so-called digital explosion. The concomitant rise in internet users, online journals and the growth of universities offering courses in the multidisciplinary areas of information and computer science are described.

Subsection 2.4.3 introduces the literature on research abstracts in general and more specifically scientific research abstracts. The reader is guided through the multitude of definitions and descriptions of this relatively short yet highly differentiated genre.

The taxonomy of abstracts is described in Subsection 2.4.4. Seven fields are introduced by which abstracts can be classified, resulting in a myriad of potential permutations. Despite the commonly perceived narrowness of this genre, a substantial amount of diversity can be discovered once researchers venture out of their own specialism.

2.4.2 Importance of scientific research abstracts

The genre of English scientific research abstracts is of importance for three reasons. First, English is set to remain as the dominant language of science. Second, the meteoric rise in the number of online publications makes reading all the literature on a topic an impossibility thereby increasing the importance of research abstracts. For writers (and publishers) the aim is to secure readers for the full article while for readers the aim is to filter the full articles to read. Third, scientists need to write abstracts in English for their research articles regardless of the language of the accompanying article. For many aiming to further their careers, it is necessary to publish in English (Reich, 2013; Montgomery, 2013).

English as the language of science

The transition away from Latin to English as the learned language of scholarly publication in the sciences in Europe occurred from the 17th century to the end of the 19th century. Since then, English has become the *de facto* language for scientific articles (Englander, 2013, p.3; Ammon, 2011; Simionescu and Simion, 2004). English is likely to remain *in situ* in the "foreseeable future" (Graddol, 1997), given that the number of users of English worldwide in 2019 is estimated at 1.268 billion, two-thirds of whom use English as a second language (Eberhard, Simons, and Fennig, 2019) and the research journal ranking bias towards English (Lillis and Curry, 2010, p.18).

English is therefore highly likely to continue to be the dominant language in science and technological research in the foreseeable future (Graddol, 1997; Hyland, 2006, p.24) with "scholars in many countries seek[ing] to publish their 'best in the West'" (ibid., p.5), maintaining the hegemony of English in the academic world dominated by Anglo-American culture (ibid., p.151). Cross and Oppenheim (2006, p.429), on a more positive note, conclude that a common *lingua franca* facilitates greater access. More recently, Hyland (2010, p.83) claims that "English is now unquestionably the language of international scholarship and an important medium of research communication for non-native English speaking academics around the world".

Increase in online publications

The growth rate of journals (Okulicz-Kozaryn, 2013) and increased online accessibility to abstracts add to the importance of scientific research abstracts. The move from bound journals to electronic publication means that research abstracts are now easily found using search engines. Google Scholar is a particularly powerful search engine enabling users to find abstracts through any net-enabled device. There has been an exponential increase in number of abstracts for conference proceedings and scientific journals that are available online. In 2017, an estimated 7.8 million researchers were engaged in the fields of science and engineering according to the *UNESCO Science Report* (UNESCO, 2018, p.14). All open-access journals and most pay-for-access journals publish research abstracts online, making them freely available to not only the scientific community, but anyone with access to the internet. Given the proliferation of scientific research abstracts and the vast size of the potential readership of any one abstract, writing persuasive abstracts is of even more importance to attract readers. There are now more journals than ever before. There are 22,000 online journals that are Scopus-indexed, housing 69 million core records with around 3 million publication records being added each year (R. Johnson, Watkinson, and Mabe, 2018, p.26). The number of online journals is increasing exponentially and given the lack of size constraints per electronic issue, the number of research articles may also rise. "The overall growth rate for scientific publication over the last few decades has been at least 4.7% per year, which means doubling publication volume every 15 years" (Okulicz-Kozaryn, 2013, p.679).

The improvement in search engine algorithms allows easier access to research abstracts in many scientific domains. Most, but not all, pay-for-access journals allow non-subscribers to view the abstracts of research articles. Additionally, since web crawling bots, such as Googlebot access the full text of abstracts search engine keyword ranking can generate results based on the actual wording in the abstracts whereas previously keywords were selected by the author or publisher. This alone results in a rise in the number of abstracts found as compared to accessing online journals in the last millennia.

There is a phenomenal increase in number of scientific publications, the availability of research abstracts online, and the volume of research output. This deluge of

information has made comprehensive literature reviews for any scientist an onerous task. In order to wade through the abstracts, researchers have to set appropriate search parameters and rely on search algorithms to find suitable hits. In the current era in which prestigious journals tend to be controlled by commercial publishers, such as Taylor and Francis, Elsevier, and Springer Science, abstracts play a prominent and important role. For the reader, abstracts provide a time-efficient way to get the key information of a research article and make the decision to read the article or continue the literature search (Cross and Oppenheim, 2006, p.429). For the publisher, they act as tasters to entice the reader to pay for access via either individual or institutional subscriptions.

Importance to academics

> *In science, the credit goes to the man who convinces the world, not to the man to whom the idea first occurs.*
> - Sir William Osler

In the words of Hyland (2004, p.63) research abstracts are a "genre critical to disciplinary knowledge-making and therefore to the work of academics". The vast majority of scholarly journals require the author(s) to submit an abstract in addition to research article (Martıín, 2003). Latour and Woolgar (1979, cited in Hyland, 2004, p.3) argue that scientists in research laboratories spend more time and energy on producing research papers (including their abstracts) than on making discoveries. The importance of scientific research abstracts is shown by exploring their functionality, the personal motivations for writing them and the potential professional benefits to authors.

In order to enable researchers to disseminate their work to potential readers an effective abstract is essential. Of the many researchers that read abstracts, only a minority will continue to read at least some part of the full research article (Swales, 1990, p.179). To quote Hoffmann (2010, p.312), "virtually all of a scientist's work will be judged first (and often last) based on an abstract". Bloor (1984 as cited in Swales, 1990, p.179) noted that Spanish academics were often required to submit English abstracts for research articles published in Spanish. For scientists working in academia, not only are abstracts a summary of the research undertaken, but they are also an essential element in the pursuit of publishing in prestigious journals and presenting at international conferences, both of which enhance the researcher's curriculum vitae. This, in turn, increases opportunities for tenure, promotion and securing funding. Researchers who make important scientific contributions are those who gain the most recognition (Garvey, 1979, p.2; Di Bitetti and Ferreras, 2017), and to ensure that the recognition is received, an effective research abstract is essential to secure publication and then reader attention. Ventola (1994, p.333) describes the function of English abstracts to circulate reported results worldwide. In short, researchers are likely to be instrumentally motivated to draft abstracts to increase

not only the readership of their research, but also to support the "cycle of credibility" (Latour and Woolgar, 1986, p.200), and accrue potential benefits (Hayer et al., 2013, p.352) to both their professional and financial standing. The phrase "publish or perish" to scientists in academia has come to mean "publish *in English* or perish" (Ventola, 1992, p.191; Lillis and Curry, 2010, p.1).

2.4.3 Definition and function of abstracts

Genre of research abstracts is just one of the nodes in the interconnected intertwined "constellation" (Swales, 2004, p.12) or "genre colony" (Bhatia, 2004, pp.65–96) of academic discourse, which Hyland, Huat, and Handford (2012, p.4) describe as "more of a myriad of texts differing across contexts than a single monolithic entity". Within the genre chain (Fairclough, 2003, p.34; Swales, 2004), research abstracts are a high stakes genre that can affect the reputation and career of a scientist. Research abstracts do not need to be read in conjunction with the full research document, and so can function as independent discourses (Van Dijk, 1980) or what Bazerman (1984, as cited in Swales, 1990, p.179) describes as a detached status of representation. An abstract is essentially a summary of a research document and can be considered as a genre itself. However, other researchers do not consider them to be a genre by themselves, but simply one part of a genre. To go some way to resolve this semantic argument, Ayers in an unpublished dissertation (1993, cited in Dudley-Evans, 2000) coined the term *part-genre*. Martıín (2003) asserts that fulfilling a specific function and having specific cultural properties, scientific research abstracts should be considered as a genre.

Hyland (2002, p.186) writes:

> Abstracts are one the most studied genres of the academy, their brevity and clear purpose making them ideal for genre studies. Several researchers have noted their value as a vehicle for projecting news value and promoting the accompanying article by encouraging the reader to continue into the main paper. This is typically done by a structure which foregrounds important information for easy access and grammatical features which highlight the novelty and immediacy.

The purpose of abstracts is claimed to be to provide a summary of the research article. But, arguably the desire to entice readers to read the full article provides a higher purpose of promoting the article through "selective representation" (Hyland, 2004, p.64) to encourage the reader to read the accompanying article. Research abstracts have been described and defined by numerous researchers over the last few decades. Research abstracts may be considered a "representation" (Bazerman, 1984, p.58 cited in Swales, 1990, p.179) or more specifically "an abbreviated, accurate representation" (ANSI, 1979, p.1 cited in Bhatia, 1993, p.78). Research abstracts may be considered a "summary" (Kaplan *et al.*, 1994, p.405 as cited in Hyland, 2009, p.70)

or again more specifically a "concise summary" (Lorés, 2004, p.281) or a "factual description or summary" (Bhatia, 1993, p.78), or a "distillation" (Swales, 1990, p.179) or "crystallization" (Salager-Meyer, 1990, p.367). Salager-Meyer (ibid.) focuses on the reduction of information available in abstracts compared to their associated articles. Rather straightforwardly, Swales (1990, p.179) refers to them as both *front matter* and *summary matter*. The role of abstracts can be used to define them as a useful filtering device to manage the ever-increasing information flow (Ventola, 1994, p.333) or as a gateway into the research literature (Hartley and Benjamin, 1998) given their role in persuading readers to continue reading. Based on a Key Word In Context (KWIC) analysis of promotional vocabulary (e.g. *importance, interest*), scientific abstracts tend to emphasize the novelty and benefit of their research (Hyland, 2004, p.76). To quote Hyland (2012b, p.33),

> Novelty has to be sold to peers as a valid contribution, and this is most obviously achieved by establishing explicit intertextual links to existing knowledge through citation and the marketing of the newsworthy in the structure of research papers. It is apparent, for instance, in abstracts where writers seek to gain readers' attention and selectively highlight what they are likely to find new.

In the words of Hyland (2004, p.63), research abstracts are a "genre critical to disciplinary knowledge-making and therefore to the work of academics". Research abstracts have multiple functions. Huckin (2001) points out their function as:

1. stand-alone mini-texts that summarize the topic, methodology and findings;
2. their function as screening devices that help readers decide to read the full text; and
3. their role as previews, or advance organisers, to help readers create a schema to follow the whole tex

This priming predisposes readers by creating a schema for the article. In the same vein, Cross and Oppenheim (2006, p.429) note that abstracts can be viewed as providing language preparation for reading the whole article, which may be described as serve a priming function (Hoey, 2005, p.13) and in doing so aids comprehension of the article (Cross and Oppenheim, 2006, p.429; Graetz, 1982).

In order to enable researchers to disseminate their work to potential readers an effective title is essential (Anthony, 2001; Haggan, 2004; Hartley, 2007; Jamali and Nikzad, 2011; Sagi and Yechiam, 2008; Soler, 2011; Y. Wang and Bai, 2007) since no readers would find the research article otherwise. The importance of the abstract then comes into play to persuade the reader of the importance, novelty and substance of their result. Axiomatically, more researchers and students read the titles of research abstracts than the accompanying abstracts. Of those that move on to read the abstract, only a minority will continue to read at least some part of the full research article

(Swales, 1990, p.179). To quote Hyland (2004, p.86), "[t]itles and abstracts are written to announce major findings and interpretations and to foreground what is innovative in the work". Hyland (ibid., p.63) also states:

> After the title, the abstract is generally the readers' first encounter with a text, and is often the point at which they decide to continue reading and give the accompanying article further attention, or to ignore it. The research and the writer are therefore under close scrutiny in abstracts and, because of this, writers have carefully, and increasingly, tended to foreground their main claims and present themselves as competent community members.

2.4.4 Types of abstracts

> *Taxonomy is described sometimes as a science and sometimes as an art, but really it's a battleground*
> - Bill Bryson, A Short History of Nearly Everything

Research abstracts can be categorized in numerous ways. Table 2.3 shows the range of fields in which variation can occur and the types of options available in each field. Some fields such as length, content and sequence are presented as dichotomies; but, in reality, each of the two options are more like end points on a cline. Each of the fields shown in 2.3 are described in turn.

TABLE 2.3: Types of scientific research abstracts

Field	Options
Publication	Journal, Conference, Dissertation
Length	Regular, Extended
Content	Informative, Indicative
Type	Regular, Structured, Graphical
Audience	Specialist, Generalist, Lay
Sequence	*a priori*, *post hoc*
Status	Submitted, Accepted, Rejected
Discipline	Linguistics, Medicine, Image processing, etc.

Field 1: Publication

Categorizing abstracts according to their publication (journal, conference or dissertation) is also an oversimplification. Swales and Feak (2009) in their authoritative book categorize abstracts into four broad categories based on the type of research document, namely research article, short communications, conference and dissertation. Conference abstracts include those submitted as proposals to be vetted and those that are published in the conference handbook and/or the proceedings. In some domains there is little difference between journal and conference abstracts, particularly when full papers must be submitted prior to the conference. A sub-division of conference

abstracts are promissory abstracts that are used to secure conference presentations in the hope that the research will have come to fruition. Swales and Feak (ibid., p.43) consider conference and journal abstracts to be different genres since their contexts and purposes are different with conference abstracts being "independent, free-standing texts" and journal abstracts part of a whole possibly prefixed with the title "abstract". In terms of purpose, conference abstracts are more promotional with the aim of either persuading the reviewer to accept the presentation or participants to attend. Samar et al. (2014) describe how moves and steps in applied linguistics conference abstracts are used to "sell" or persuade readers, and how the use of these moves and steps vary with research (e.g. type) and researchers (e.g. novice). Dissertation abstracts include abstracts for all university degrees whether the abstracts are connected to capstone projects, graduation theses, master or doctoral theses or dissertations.

Field 2: Length

There is no universal agreement on the length of an abstract, and in fact the length specifications advocated by journal editors frequently do not match the actual abstracts published in the journals. However, as a rule of thumb, most research abstracts are between 200 and 500 words. Within a discipline regular abstracts are shorter than extended abstracts. However, the disciplinary variation is so great that extended abstracts in one discipline may be shorter than regular abstracts in another discipline. Extended abstracts tend to range from around 500 to 1500 words depending on the research discipline. In some journals, research letters, research notes or short communications are a similar length to extended abstracts.

Field 3: Content

The dichotomy of informative (complete) or descriptive (indicative) abstracts was the focus of attention in earlier research on abstracts (Day and Gastel, 2006; Graetz, 1982; Jordan, 1991; Lorés, 2004). Informative abstracts, according to Jordan (1991, p.507), "act as a report in miniature". Indicative abstracts, however, do not provide many specific details and tend to only point out key points, such as the subject area and main findings.

Field 4: Type

For the first 250 years of the three-century history of scientific publishing, abstracts were not part of the format of a research article. Traditional abstracts became commonplace from the 1960s (Swales and Feak, 2009). There is also a trend since the 1980s, particularly in the field of medicine, towards structured abstracts which have specified subsections (Hartley, 2003). Structured abstracts were designed to help time-pressured medical practitioners extract the main details for abstracts more efficiently. With the flexibility of digital publishing, in some scientific fields such as

organic chemistry, graphical abstracts (Hayashi, Tomioka, and Yonemitsu, 1998; Abdelmoneim, 2010; Lane, Karatsolis, and Bui, 2015; Yoon and Chung, 2017) depicting the main finding visually, have gained in popularity.

Field 5: Audience

Abstracts are in general pitched to specialists. No concession is made to readers without a background in that discipline. Although the lay reader can get the gist of the message, almost all of the specific details are beyond their comprehension due to an overload of new terms. For example, in information theory terms like *bit error rate, GMSK, block, signal envelope* and *Hamming code* are unlikely to be understood by lay readers. In some wide-ranging disciplines, abstracts need to be written so that generalists within the same broad discipline can understand the main thrust of the research. However, some abstracts, such as those used in multidisciplinary popular journals, such as *Nature* and *Science*, need to be written for educated lay readers.

Field 6: Sequence

Journal abstracts are anecdotally written *ex post facto*, (Swales, 1990, p.181; Groves and Abbasi, 2004, p.470). Some specialist informants, however, noted that they draft envisaged abstracts prior to starting a new research project. These abstracts may go through numerous rounds of revisions with the result that an abstract written *a priori* and the *post hoc* abstract of the same research project may radically differ.

Field 7: Status

Abstracts may be categorized based on their status, that is whether the abstract (and its accompanying article) was submitted, accepted or rejected. Koyamada et al. (2018) analyzed abstracts of papers submitted for publication in the *Journal of Visualization* and found that accepted abstracts contained more RESULT MOVES and DISCUSSION MOVES than rejected abstracts.

Field 8: Discipline

There are numerous ways of categorizing disciplines. Becher and Trowler (2001, p.28) suggest a taxonomy classifying disciplines into four domains based on two dimensions, as shown in Table 2.4.

TABLE 2.4: Classification matrix for scientific disciplines

	Hard sciences	**Soft sciences**
Pure sciences	e.g. physics	e.g. anthropology
Applied sciences	e.g. mechanical engineering	e.g. education, law

Source: Becher and Trowler (2001)

Becher and Trowler (ibid.) also note that disciplines are being fragmented into sub-disciplines. Berry (1981, p.400) describes the creation new styles of journals by drawing on the structure and linguistics features of Ecclesiastes XII, the twelve chapter of the Book of Ecclesiastes:

> To the making of many books there is no end.— Ecclesiastes XII, V.12. In the beginning was the General Scientific Journal. And the General Scientific Journal begat the Specialty Journal, and the Specialty Journal begat the Subspecialty Journal. And the Subspecialty Journal begat the Single-Subject Journal, whether according to class of compound, specific disease, or methodology. And the Single-Subject Journal begat the Interdisciplinary Journal to link up the specialties separated at an earlier evolutionary date. And the scientific community saw that the journals were good, and they were fruitful and multiplied.

This disciplinary fragmentation may be considered indicative of greater specialization, but as Balietti, Mäs, and Helbing (2015) showed fragmentation hinders research progress, which is in line with Kuhn's view of pre-paradigmatic science (Kuhn, 2012).

2.4.5 Section summary

This section has argued that scientific research abstracts are a genre worthy of further research based on their considerable importance to both readers and writers. Readers benefit by increased productivity as abstracts are a time-efficient way of ruling out research articles that are not central to their own research agenda. Writers benefit by the career advancement and the profiling-raising nature of being published in top-tier journals. Both the associated article and the abstract are important, but as a genre of persuasion, the abstract provides that all-important first opportunity to "sell" an idea to the reader.

As has been shown, English is set to remain the language of science in the foreseeable future. As the literature grows exponentially, abstracts are becoming increasingly more important to enable readers to eliminate articles that are less relevant to their own research study. With advances in natural language processing, it may well be that automated abstract selection starts to take over the role, but until that time narrowing down the literature to read remains a time-consuming task for any researcher.

The genre of research abstracts is somewhat fragmented with a variety of named typologies including the recent additions of structured and graphical abstracts; but even within a particular typology, there is no broad agreement of what moves should be included and in what sequence the moves should be presented.

This section has established that further research is deserved to better understand the functions, features and format of research abstracts. What appears to be lacking is a comprehensive framework that provide a way to map both idealized abstracts and real-world artefacts in terms of features, functions and format. In the following

section, (Section 2.5), a state-of-the-art survey of the published literature on rhetorical organization in scientific research abstracts is provided.

2.5 Rhetorical organization of scientific research abstracts

> *Academia is ruled by a pathological herd mentality. Stick out, and be prepared to get ostracized.*
> - Gad Saad

2.5.1 Overview

This section critically reviews the literature on rhetorical organization of research abstracts, focusing on studies that deal with scientific disciplines.

Subsection 2.5.2 discusses the selection of moves and explains the pedagogic benefits for opting for a set of moves that can help novice writers understand how abstracts are structured. In rhetorical organization, the sequencing of the moves is paramount. Sequence can be described mathematically, and so a brief introduction to the concepts of combinations, permutations and factorials is provided.

The following subsection, (Subsection 2.5.3), presents an overview of the dominant existing models of rhetorical structures of research abstracts. Swales' three-move CARS model, the traditional four-move IMRD model, the five-move models of Santos and Hyland, and the more recent six-move model for conference abstracts of rhetorical organization are described, discussed and evaluated.

Subsection 2.5.4 identifies the deficiencies in the current models of rhetorical organization, highlighting a gap in the research literature. The deficiencies discussed relate to the lack of validity of the models based on their lack of coverage. The models appear specific to disciplines or clusters of disciplines. Another issue concerns the lack of double annotation and the lack of reported inter-annotator agreement. Although it may well be the case that a single proficient annotator can more accurately annotate than a group, the scientific method is founded on replicability and if the results cannot be reproduced, the validity of the results may be criticized.

2.5.2 Moves and sequences

There are two elements to rhetorical organization, namely the presence or absence of particular rhetorical moves and the sequence of those moves. The moves frequently used in the extant models (as described in Subsection 2.3.2) are INTRODUCTION, BACKGROUND, GAP, PURPOSE, METHOD, RESULTS and DISCUSSION. If an abstract contains only one move, there is no sequence. However, single-move abstracts are rare. In most abstracts two or more moves are used, and so the author needs to choose the appropriate sequence. Let us consider some factors that authors need to consider when ordering two moves: MOVE ONE and MOVE TWO. Possible factors include:

1. Importance

2. Novelty
3. Substance
4. Rigour
5. Givenness
6. Length

We know that reviewers are time-pressured and may reject submissions simply based on reading abstracts (Groves and Abbasi, 2004). However, given that abstracts are a form of persuasion (Hyland, 2002; Hyland, 2004; Samar et al., 2014), convincing readers of the importance, novelty, substance and rigour of the research ought to be of primary concern. Both rigour and novelty are highly valued in many disciplines, including mathematics (Burton, 2004) and biology (Brembs, 2019). Other discourse communities may value these concepts differently.

The importance of the research is often shown in the RESULT or the DISCUSSION moves. The novelty may be shown in the INTRODUCTION, METHOD or RESULT MOVES. Substance and rigour are likely to be shown in the METHOD MOVE. This gives the author a number of factors to consider when sequencing moves. Consider the specific example of sequencing two moves: METHOD and RESULT.

1. Sequence 1: METHOD followed by RESULT
2. Sequence 2: RESULT followed by METHOD

If the author aims to emphasize the novelty, rigour or substance of the method, *Sequence 1* may be the most appropriate while *Sequence 2* may be the most appropriate choice when the author aims to emphasize the result.

Generic integrity also needs to be considered since ordering moves in an unexpected way may be perceived negatively by the audience, and so following precedents might be the best option.

In addition to the focus on the impact that the author intends the abstract to make on the reader, there are grammatical principles at play. For native speakers, these principles are second nature, but to users of English as an Additional Language, these principles may be learned explicitly or acquired through extensive exposure to language. The two core principles are givenness and end weight. Typically, within a sentence, given or known information is placed before new information. In addition, substantially longer expressions tend to be placed after shorter expressions. When either of these principles are broken, then an information focus is created (Biber, Johansson, et al., 1999; Blake, 2015a).

As can be seen, there are many micro-decisions to be made in the sequencing of just two moves. Axiomatically, the number of micro-decisions increases exponentially as the number of moves increases.

Combinations, permutations and factorials

A number of researchers have proposed particular combinations or permutations of moves. A combination of moves refers to the particular moves that occur in an abstract without specifying the order of the moves. A permutation is used to refer to a particular sequence of particular moves. For example, for a two-move abstract using (move x and move y), there is one combination (x and y) but two permutations (x,y and y,x). As the number of moves increases, the number of combinations increases linearly; but the number of permutations increases exponentially. A factorial is the product of an integer and all the integers below it. The factorial of three is (3!) is six, the factorial of four (4!) is 24 and the factorial of (5!) is 120. Factorials can be used to calculate the number of potential permutations of a move structure from the number of moves, assuming that each move is used once, i.e. there are no omissions or repetition. Figure 2.5 shows all the possible permutations for three moves.

TABLE 2.5: Possible permutations of three rhetorical moves

Number	MOVE ONE	MOVE TWO	MOVE THREE
1	Introduction	Method	Result
2	Introduction	Result	Method
3	Method	Introduction	Result
4	Result	Introduction	Method
5	Method	Result	Introduction
6	Result	Method	Introduction

The number of permutations increases significantly if omission or repetition are permitted. For example a five-move model which allows the omission or addition of one move (but not the addition of one move and the omission of a different move) can be calculated by ($5! + 4! + 6!$), giving 864 possible permutations.

2.5.3 Current models of rhetorical organization

> *Knowledge about the process being modeled starts fairly low, then increases as understanding is obtained and tapers off to a high value at the end.*
> - Thomas Samuel Kuhn, The Structure of Scientific Revolutions, 1962, p. 46

Lau (2004) analysed the move structure of a corpus of 80 life science abstracts, 50 abstracts written by Taiwanese Ph.D. candidates with 30 abstracts by foreign scholars, and showed that while the abstracts written by foreign scholars contained four or five moves with the METHOD MOVE frequently being omitted, half of the abstracts written by Taiwanese omitted the INTRODUCTION, PURPOSE and METHOD MOVES. This is a clear example of necessity of raising the awareness of obligatory and optional moves to novice writers.

The research on rhetorical organization of research abstracts could be broadly divided into two categories: studies that propose rhetorical structures and studies

that use proposed rhetorical structures. Those studies that propose structures almost invariably define their own moves and then propose a particular sequence. In this subsection the proposed models are described and evaluated.

One of the earliest published studies on rhetorical organization of research abstracts was conducted by Nwogu (1990) who discovered that the rhetorical structure of a medical abstract may not represent that of its related article. In this study, he found that out of eleven moves identified in medical research abstracts, two moves were obligatory while nine moves were optional. Numerous researchers (Anderson and Maclean, 1997; Abdelmoneim, 2010; Day and Gastel, 2006; Holmes, 1997; Holtz, 2011; Hyland, 2004; Jordan, 1991; Lorés, 2004; Pho, 2008; Salager-Meyer, 1992; dos Santos, 1996; Tseng, 2011; Tu and S. Wang, 2013) have studied the rhetorical structure of abstracts in various academic disciplines. However, none of these studies dealt with disciplines with applied sciences, such as the multidisciplinary field of information science, which includes disciplines such as wireless communication, information theory and image processing. Proposed rhetorical move structures often provide the option for some moves to be obligatory while others are optional (see, for example, Huckin, 2001; Nwogu, 1990; Pho, 2008). This helps a proposed model of move structure fit a wide variety of abstracts. In fact, a vaguely defined move structure is often easily fit over an abstract, but the vagueness leads to problems in the interpretation. Once scientists in Europe adopted the IMRD structure (Sollaci and Pereira, 2004), the genre of the research article was born. However, the rigid format has critics. Lock (1988, p.156) described the IMRD model as an "intellectual straitjacket".

A chronological list of the published studies proposing or using different models of the rhetorical organization of research abstracts is given in Table 2.6. Ten of the eleven models were either based on a corpus with a stated size of fewer than 100 abstracts or no corpus at all. Two of the eleven models used corpora with abstracts from various disciplines (n = 77 and n = 800). The only model that was based on a relatively large corpus comprising abstracts from various fields was Hyland (2004). This corpus comprised abstracts from both the sciences and humanities. The METHOD MOVE and RESULT MOVE occur in eight models although Hyland (ibid.) uses the PRODUCT MOVE to show that some results are artefacts. The DISCUSSION MOVE is also mentioned in eight models although the naming of the move varies with DISCUSSION MOVE, SUMMARY MOVE and CONCLUSION MOVE being used. The models based on single disciplines may reflect that discipline, but given the radical differences between the proposed models for medical, applied linguistics and pyschology; these models are unlikely to fit all disciplines. Apart from the frequency of the inclusion of the METHOD MOVE, RESULT MOVE, and DISCUSSION MOVE; there appears to be little agreement on the naming of the moves, the demarcation of moves (e.g. between summary, discussion and conclusion), and the order of the moves. What is lacking is an evidence-based model that can describe how rhetorical moves are structured in all scientific disciplines.

TABLE 2.6: Studies proposing or using different models of rhetorical organization of research abstracts in chronological order

Year	Surname	Move structure	Short form	Field	Corpus size
1985	Graetz	Problem, Method, Result, Conclusion	PMRC	Various	87
1990	Swales	Introduction, Method, Result, Discussion	IMRD	n/a	n/a
1990	Swales	Territory, Establish and Occupy niche	CARS[a]	n/a	48
1990	Salager-Meyer	Statement of problem, Purpose, Methods, Results, Conclusion, Recommendation	SPMRCR	Medical	77
1993	Bhatia	Purpose, Method, Result, Conclusion	PMRC	n/a	n/a
1994	Ventola	Combinatory structure of informative-indicative	n/a	n/a	
1996	Santos	Situating, Presenting, Describing method, Summarizing results, Discussing	SPMRD	App Ling[b]	94
1998	Hartley & Benjamin	Background, Aims, Methods, Results, Conclusions	BAMRC	Psych[f]	60
1998	Yakhontova	Outline, Justification, Introduction, Summary, Highlight	OJISH	App Ling[b]	45
2002	Samraj	Territory, Establish and Occupy niche	CARSmod	WilBeh[d] & ConBio[e]	12
2004	Hyland	Intro, Purpose, Method, Product, Conclusion	IPMPrC	Various	800
2009	Swales & Feak	Outlining, Justifying, Method, Summarizing, Highlighting, Further observations	OJMSHF	unknown	unknown

[a] Create-a-research-space, [b] Applied Linguistics, [c] Conference abstracts, [d] Wildlife behaviour, [e] Conservation biology, [f] Psychology

When research abstracts became part of the research article, they were expected to reflect the organization of the research article. Bhatia (1993, pp.78–9) describes the four-step move structure of abstracts which reflects the structure of the paper overall as comprising PURPOSE or INTRODUCTION, METHOD, RESULTS and CONCLUSION or DISCUSSION. The IMRD structure, however, was at odds with the early findings in medical research articles investigated by Nwogu (1990). In the field of scientific abstracts, IMRD or variations of it predominate in textbooks aimed at helping researchers write scientific articles (Hoffmann, 2010; Swales and Feak, 2009). Graetz (1982) in her study of abstracts in various disciplined renamed INTRODUCTION to PROBLEM and DISCUSSION to CONCLUSION.

Three-move model

Swales (1990) describes rhetorical moves and steps in the Create-A-Research Space (CARS) model for introductions in research articles. This model, shown in Table 2.7, has since been widely applied by other researchers to research abstracts.

TABLE 2.7: Create-A-Research Space (CARS) model

Moves	Steps
Move 1 Establishing territory	Step 1 Claiming centrality *and/or*
	Step 2 Making topic generalization(s) *and/or*
	Step 3 Reviewing items of previous research
Move 2 Establishing a niche	Step 1A Counter-claiming or
	Step 1B Indicating a gap or
	Step 1C Question-raising or
	Step 1D Continuing a tradition or
Move 3 Occupying the niche	Step 1A Outling purposes or
	Step 1B Announcing present research
	Step 2 Announcing principal findings
	Step 3 Indicating article structure

Source: Swales (1990, p. 141)

The CARS move structure has come under scrutiny based on its lack of clear criteria on what constitutes a particular move or step, lack of examples and lack of objectivity (Anthony, 1999; Crookes, 1984; Dudley-Evans, 2002; Lorés, 2004). The model was accepted initially by many in the humanities as it appeared to match their reality and so appeared to fall victim to confirmation bias.

Four-move model

The four-move IMRD structure has often been harnessed by numerous researchers in their studies of abstracts (Bhatia, 1993, pp.78–79; Salager-Meyer, 1990; Salager-Meyer, 1992; dos Santos, 1996). When the IMRD structure did not fit particular disciplines, researchers incorporated features of both IRMD and CARS, creating combinatory structure abstracts, (Lorés, 2004, pp.283–6; Ventola, 1994, p.335). Numerous researchers have proposed variations or modifications to the IMRD structure (Abdelmoneim, 2010; Graetz, 1982; dos Santos, 1996; Pho, 2008). Lorés (2004, p.283)

found that the IMRD move structure was a better fit for around 60% while CARS move structure was a better fit for around 30% of her corpus of 36 research abstracts in linguistics. Given the necessity to create new combinatory structural models, it is clear that neither CARS nor IMRD models are suitable as descriptive models as they cannot cover all cases.

Five-move model

dos Santos (1996, pp.484–490) proposed that a fifth move SITUATING-THE-RESEARCH MOVE precedes the four moves suggested by (Bhatia, 1993). This move is achieved by describing the current state of understanding of the research area, citing previous research, extending previous research or stating a problem.

Hyland (2004, p.67) introduced an alternative five-move structure for abstracts consisting of INTRODUCTION, PURPOSE, METHOD, PRODUCT and CONCLUSION MOVES as shown in 2.8. In this five-move structure, PURPOSE has been elevated from the sub-move to move status. The renaming of RESULT to PRODUCT is aimed at more easily incorporating research that aims to create theories, artefacts and algorithms rather than test hypotheses.

TABLE 2.8: Classification of rhetorical moves in research article abstracts

Move	Function
Introduction	Establishes context of the paper and motivates the research or discussion.
Purpose	Indicates purpose, thesis or hypothesis, outlines the intention behind the paper.
Method	Provides information on design, procedures, assumptions, approach, data, etc.
Product	States main findings or results, the argument or what was accomplished.
Conclusion	Interprets or extends results beyond scope of paper, draws inferences, points to applications or wider implications.

Source: Table 4.1 in Hyland (2004, p.67)

In his analysis of 800 abstracts, fewer than 5% of abstracts contained all five moves, but 94% included the product move (ibid., pp.68–69). Hyland (ibid., p.70) also notes that in the hard sciences approximately 60% contained the METHOD MOVE while around 30% contained the INTRODUCTION MOVE. He found that abstracts in the humanities dedicate more importance to the INTRODUCTION MOVE due to the necessity to discuss or define issues as opposed to establishing empirical truths (ibid., p.72).

In disciplines in which the research landscape changes quickly such as in some areas of information science and computer science, there is little difference between conference abstracts and journal abstracts. Most likely, as most of conferences in this discipline require the submission of a full research article for acceptance to the conference in contrast to the humanities and softer sciences in which an abstract is usually sufficient to gain acceptance. Yakhontova (1999) identified a five-part structure (see Table 2.9), which is rather different to the other rhetorical structures.

The first three moves may be broadly grouped into an INTRODUCTION MOVE using the four-move IMRD model while the fifth move could be classed as the RESULT MOVE and/or DISCUSSION MOVE.

TABLE 2.9: Five-part move structure for conference abstracts

Move	Description
1	Outlining the research field
2	Justifying this particular piece of research/study
3	Introducing the paper to be presented
4	Summarizing the paper
5	Highlighting its outcome/results

Six-move model

Swales and Feak (2009, p.45) more recently promote the use of a six-part move structure for conference abstracts, noting that not all the moves are obligatory, as shown in Table 2.10. In this structure moves 1 and 2 would come under the INTRODUCTION MOVE in the IMRD model and moves 4 and 5 could be classed as RESULT MOVE In short, this six-part move structure may be considered a more finely grained version of the IMRD model.

In some scientific disciplines conference proceedings are more prestigious than others. This is because of the fast-changing nature of their discipline and the need to disseminate research quickly. It should also be noted that, in general, conference acceptance in the humanities is based on the conference abstract alone while in the sciences acceptance is based on the complete research article. Some scientific conferences require prior submission of a conference abstract, but this is to enable the organisers to select suitable reviewers rather than to assess the content of the abstract.

TABLE 2.10: Six-part move structure for conference abstracts

Move	Description
1	Outlining/promoting/problematizing the research field or topic
2	Justifying this particular piece of research/study
3	Methodological, demographic, or procedural comments
4	Summarizing the main findings
5	Highlighting its outcome/results
6	Further observations (implications, limitations, future developments)

The five- and six-move models provide researchers with more choices on how to annotate moves, and so should be able to better fit the moves. However, published abstracts in various disciplines appear to frequently diverge from the expected sequence in the models described. These models focus on obligatory and optional moves, but do not focus on sequence. What is needed is a model that can catch the complexities occurring in move sequences, such as expected adjacency pairs of moves, fronting of moves and reuse of moves or move sequences within abstracts.

TABLE 2.11: Studies investigating rhetorical organization of research abstracts

Year	Surname	Relevant finding	Discipline	Corpus size
1990	Nwogu	Abstract may not represent article structure	Medical	45
1995	Berkenkotter & Huckin	Feature and format changes increase news value	Various	350
1996	Santos	Embedding of moves within sentences	App Ling[b]	94
1997	Anderson & McClean	Embedding of moves within sentences	Medical	80
2001	Huckin	Omission of purpose move	Biomed[c]	unknown
2003	Hartley	Structured vs. unstructured	Pysch[d]	24
2004	Lorés	Frequency of different structures (IMRD, CARS, combination)	Ling[e]	36
2004	Lau	Non-native English Speakers omit moves	Life[f]	80
2006	Cross & Oppenheim	Generic structure: themes	ProZoo[g]	12
2008	Pho	3 obligatory moves	App Ling[b] & EdTec[i]	30
2009	Holtz	Comparison to full article	Various	94
2011	Cava	Recurrent expressions	Various	1035
2011	Tseng	Four moves are most frequent	App Ling[b]	90
2013	Dorós	Fewer moves correlates less clarity	Ling[e] & Lit[j]	40
2013	Suntara & Usaha	Variation in move order between two related disciplines	App Ling[b] & Ling[e]	200

[a] Popular science, [b] Applied Linguistics, [c] Biomedical, [d] Pyschology, [e] Linguistics, [f] Life sciences (biological sciences), [g] Protozoology, [h] Computer-assisted language learning, [i] Educational technology, [j] Literature

A chronological list of studies investigating rhetorical organization of research abstracts is given in Table 2.11. This table shows how the research terrority has changed over time. The initial research by Nwogu (1990) showed that the structure of medical abstracts did not represent the structure of the article. This result, however, is probably no longer valid, given that medical journals have adopted structured abstracts. Researchers in the late 1990s (Berkenkotter and Huckin, 1995; dos Santos, 1996) found that moves were realized at sub-sentence level via embedding in medicine and applied linguistics. No research was found reporting on the presence or absence of move embedding in other scientific disciplines, though. Research on structured abstracts (Hartley and Benjamin, 1998; Hartley, 1994; Hartley, 2003; Hartley, 2007) showed the positive effect on readability, which explains the widescale adoption of structured abstracts in medical journals. What is not discussed though is why structured abstracts have not been embraced by many other disciplines. A number of studies have investigated which moves occur in particular disciplines (Huckin, 2001; Pho, 2008; Tseng, 2011) while other researchers have focused on the effect of move structure on clarity (Doró, 2013). Despite the obvious interest in rhetorical organization, no studies that focussed on the sequencing of rhetorical moves were found. This is the research niche that this present study aims to occupy.

A chronological list of studies investigating disciplinary and authorial variation of research abstracts is given in Table 2.12. These studies on variation focussed on either (1.) the presence or abscence of rhetorical moves or (2.) the linguistic features that occur in abstracts. Two reasonably large-scale corpus studies were found (Hyland, 2004; Tu and S. Wang, 2013). The key finding relating to this study for Hyland (2004) is the broad difference between hard and soft sciences with hard sciences tending towards the IMRD structure while soft sciences were perceived to follow the CARS move structure. This may be explained by the increased need to orientate readers to the research field. Tu and S. Wang (2013) in their research on abstracts from linguistics and applied linguistics found that IMRD was more common than CARS or IPMPrC (See Table 2.6).

No study was found that focussed on the lexical realization within rhetorical moves, and axiomatically no study focussed on the lexical realization with rhetorical moves across a broad range of scientific disciplines. This is another research gap in the literature.

2.5.4 Deficiencies in current models of rhetorical organization

Validity and coverage

Although there are numerous models, each model comes with deficiencies. In this subsection, the lack of validity of models and the lack of widespread coverage are detailed. The focus on adaption of models to "fit" data by fixing the model is then discussed. Finally, for almost all these models, the details on inter-annotator agreement are sparse to none.

TABLE 2.12: Studies investigating disciplinary variation in rhetorical organization of research abstracts

Year	Surname	Relevant finding	Discipline	Corpus size
1982	Graetz	pedagogic extraction of patterns	Various	87
2003	Stotesbury	Evaluation differences	24 disciplines	300
2004	Lau	NNES lacked background/intro move	life science	80
2004	Lorés	IMRD vs CARS	Ling[c] & AppLing[d]	36
2004	Hyland	Hard sciences tend to IMRD while soft sciences tend to CARS	Various	800
2005	Samraj	Variation detected	ConBio[a] & WildBeh[b]	24
2011	Holtz	Variation across disciplines	Ling[c], CompSci[f], Mech[i], Biol[j]	94
2013	Suntara & Usaha	Variation detected	Ling[c] & AppLing[d]	200
2013	Behnam & Golpour	Greater disciplinary variation across different languages in mathematics than applied linguistics	AppLing[d] & Math[e]	80
2013	Doró	literature focussed more on context	Ling[c] and Lit [h]	40
2013	Tu & Wang	IMRD more frequent than CARS and IPMPrC	Ling[b] & AppLing[c]	1000
2014	Esfandiari	four moves plus optional move	2 domains in CompSci[f]	32
2015	Masawana, Kanamaru & Tajino	Move analysis in Eng[g] domains	Eng[g]	56

[a] Conservation biology, [b] Wildlife behaviour, [c] Linguistics, [d] Applied linguistics, [e] Mathematics, , [f] Computer science, [g] Engineering, [h] Literature, [i] Mechanical engineering, [j] Biology

Models are often proposed based on small-scale investigations. Other researchers often apply a model of a proposed move structure to different disciplines. This may be due to overly bold claims made by researchers proclaiming all-encompassing models. A model that works for one discipline does not necessarily fit another discipline. The main cause of lack of coverage of a model is its failure to account for disciplinary variation. This may be due to lack of awareness of the degree of disciplinary variation. A case in point is Can, E.Karabacak, and Qin (2016), who found in a study of 50 abstracts from the journal *English for Specific Purposes* that the expected five-move structure was frequently violated.

Martin and Rose (2012, p.15) asserted that there is an "apparent gap between the ideal genres and real word texts". This claim may hold true for scientific research abstracts. Models appear to have been devised to provide pedagogic teaching tools for ESP learners, and so is further investigation to needed to ascertain whether or not such a gap exists.

Variation in rhetorical organization has been investigated between:

- languages of abstracts (Martıín, 2003; Martıín and Burgess, 2006; Salager-Meyer, Ariza, and Zambrano, 2003);
- types of writers (Amnuai, 2019; Lau, 2004);
- different disciplines (Doró, 2013; Abdelmoneim, 2010; Hyland, 2004; Holtz, 2011; Huckin, 2001; Stotesbury, 2003; Suntara and Usaha, 2013); and
- closely related disciplines (Lorés, 2004; Samraj, 2002).

Although there are numerous studies on variation, there is still no overall framework or model that is able to encompass the rhetorical organization for every discipline.

Apart from Hyland (2004) which was based on a corpus of 400 abstracts from scientific disciplines and 400 from the humanities, the other corpora used in the other studies were limited in the number of disciplines and/or the size of the corpus.

The CARS and IMRD models have bred numerous variations and hybrids in attempts by subsequent researchers to account for non-conforming findings in small-scale studies, which are usually focussed on very narrow scientific disciplines. One such example is Samraj (2002), who proposed modifying the CARS model to account for differences in variation across disciplines of conservation and wildlife biology. The incremental alteration to models uses Occam's razor (Rasmussen and Ghahramani, 2001) to make the simplest change, but in doing so limits the chance of any radical improvement. Perhaps, what is needed is to consider a model using a starting point other than IMRD or CARS.

Issues with inter-annotator agreement

Another prevalent issue is the lack of double annotation and in the cases in which there is more than one annotator, the lack of details on inter-annotator agreement.

When inter-annotator agreement is reported, frequently only a single simple ratio of agreement is presented with no specific details on the relevant variables, such as the actual number of texts compared, the ontological units, or whether the ratio reports agreement after or before discussion of differences. In applied linguistics few studies report measures of agreement and measures of disagreement.

2.5.5 Section summary

Prescriptive advice given in the pedagogic literature tends to provide sweeping generalizations as publishers aim to market textbooks to as wide a market as possible. This is, however, a disservice to researchers working in disciplines in which their research abstracts do not conform to the standard IMRD structure. To them, there is divergence between the move structures of abstracts they read in their journals and that advocated in textbooks: a disjuncture between prescriptive advice and their descriptive reality. A current trend to data-driven learning as a way to deal with disciplinary variation may go some way to addressing this, but it is no panacea.

Most of the studies on rhetorical organization used very small corpora ($n < 100$ abstracts). Very few reported details on inter-annotator agreement, and those that did provided only sketchy details, leaving the door wide open for accusations of subjectivity and lack of replicability by proponents of the scientific method.

Of the published studies, only the study by Hyland (2004) dealt with a number of scientific disciplines. The corpus for that study, however, was rather limited in size with only 400 abstracts from the sciences. The results may or may not hold true with a larger corpus and may or may not reflect the disciplinary variation within other scientific domains.

In summary, to address the gap in the research literature, a larger-scale corpus study of a broad range of scientific research abstracts using multiple annotators is called for. The higher the inter-annotator agreement, the less room there will be critics to deny the veracity of the results. Given that currently no research provides a comprehensive picture of scientific research abstracts in terms of the rhetorical organization of moves, this could provide data that could go some way to creating such a framework.

This section has described the extant body of literature on rhetorical organization of scientific research abstracts. The lexical realization within rhetorical moves is addressed in the following section. However, in view of the fact that only the fields of linguistics and medicine have received some attention, little is known about lexical realization within rhetorical moves.

2.6 Lexical realization in scientific research abstracts

> *"When I use a word," Humpty Dumpty said, in rather a scornful tone, "it means just what I choose it to mean—neither more nor less." "The question is," said*

Alice, "whether you can make words mean so many different things."
- Lewis Carrol

2.6.1 Overview

The inextricable intertwinned relationship between lexis and grammar is addressed in Subsection 2.6.2. The lexical realization in scientific writing with a particular emphasis on scientific research abstracts is then considered. A brief description of the research on common traits in scientific writing, such as grammatical metaphors, long grammatical subjects and precision, is given.

The following subsection (Subsection 2.6.3) introduces patterns of co-occurrence and shows the relationship between collocation and colligation using a cline of co-occurrence. As there is a paucity of published research on collocation and colligation within scientific research abstracts, research from the broader field of research on scientific writing is drawn upon and discussed.

Keyness is defined and discussed in Subsection 2.6.4. The review of the literature shows that although there are numerous studies on keyness, there is little work on keyness within rhetorical moves, and no published large-scale studies on keyness within rhetorical moves in scientific abstracts or scientific writing.

Subsection 2.6.5 deals with grammatical tenses, that is the twelve pedagogic tenses that make up part of the grammatical syllabus in many courses for NNESs. Although grammar is no longer a fashionable orthodoxy in published textbooks of mainstream English Language Teaching (ELT), EAP or ESP; the importance of tenses to meaningful communication is undisputed. Without mastery of grammatical tenses, meanings are lost and confusion ensues. In this subsection, previous studies on tense within research abstracts are discussed. This review of the literature reveals a dearth of research on grammatical tenses, and none on grammatical tenses within moves in a range of scientific disciplines.

The summary in Subsection 2.6.7 draws together the lexical themes and patterns of co-occurrence and concludes by highlighting the gaps in the research literature on keyness and grammatical tenses. Given the difficulties of annotation in terms of time, resources and expertise; at the time of writing, there are no corpus studies that are able to provide insight into the relationship between moves and the lexis and grammar that are used to realize those moves in a research abstracts in a broad range of scientific disciplines.

2.6.2 Lexis and grammar

Lexis and grammar are often viewed separately by linguists, yet any grammatical feature is realized by lexis, and as such the separation of grammar from lexis is artificial. Halliday recognised the continuity between grammar and lexis and so coined the term *lexicogrammar* to encompass any aspect of the "complementary perspectives" of grammar and vocabulary (Halliday, 1991, p.32). He also asserted

that "[i]f you interrogate a system grammatically you will get grammar-like answers and if you interrogate it lexically you get lexis-like answers" (Halliday, 1992, p.64). Stubbs (1996, p.36) also states that "[t]here is no boundary between lexis and grammar: lexis and grammar are interdependent". In this research lexical will be used the broad sense acknowledging that grammatical realization in English is not possible without lexis.

Lexis and grammar of scientific writing

> *"Religion is a culture of faith; science is a culture of doubt".*
> - Richard Feynman

Science is said to have a language of its own which prescribes a particular way of looking at the world (Derewianka, 2003)). Within the scientific genres, research abstracts are particularly notable. In contrast with the humanities, abstracts in scientific disciplines are frequently impenetrable to lay readers due to the accumulation of highly specialized words, lengthy noun phrases and extensive nominalization. The difficulty of reading scientific abstracts is described by Benbrahim and Ahmad (1995, p.322):

> Abstracts ... are written for experts by experts, ... [T]he language of abstracts is terse and requires substantial background knowledge. Moreover, there would be few brave people who would argue that the abstracts of specialist documents are exemplars of plain language.

Experts in a particular field can draw upon a vast range of exotextual knowledge that is necessary to understand pithy statements and nuances of buzzwords in their specialist discipline, which most likely will be missed by those with different educational backgrounds. Biber and Gray (2010) notes that the use of structurally "compressed" forms, such as the embedding of modifiers in noun phrases, creates dense texts that expert readers can parse efficiently, but pose significant challenges to novice readers.

The following paragraphs introduce some of the known lexical features that permeate scientific research abstracts.

Grammatical metaphor Halliday (1985, p.358) coined the term *grammatical metaphor* to describe the way in which processes are repackaged as abstract concepts resulting in concise dense noun-heavy text (Biber and Gray, 2013; Biber, Conrad, and Rippen, 1998; Halliday and Martin, 1993; Hanauer and Englander, 2013). Grammatical metaphor is a "semantic category such as a process is realized by an atypical grammatical class such as a noun, instead of a verb" (Martin and Rose, 2007, p.106). Martin (1993, p.230) explains that grammatical metaphors enable more meaning to be packed into few words, therefore making scientific language more concise. Nominalisation is the "most powerful resource for creating grammatical metaphor" (Halliday, 1994,

p.352; Halliday and Matthiessen, 2004, p.656). The abstraction of a process into a nominal phrase enables writers to writers to thematize the verbal process and make it the focus on attention (Matthiessen and Halliday, 2009). The resultant reduction in the need for personal pronouns to fill the subject position results in depersonalization (Hatim and I. Mason, 1997, p.25), enhancing the perception of objectivity and allowing "the writer to give the required flavor of objectivity to his or her statements and claims" (Holes, 1995, p.260).

Lexical and information density The high frequency of nominalization (Biber and Gray, 2013) contributes to their high lexical density (Halliday and Martin, 1993) and information density (Holtz, 2009). The high lexical density of research abstracts makes them difficult to understand and write (Halliday and Martin, 1993). Holtz (2009) showed empirically that nominalization is significantly more frequent in abstracts than their accompanying research articles.

With the high lexical density of research abstracts and the assertion by researchers (Swales, 1990, p.187; Hartley, 1994, p.429; Pho, 2008, p.231) that mastery of drafting abstracts is part of the *rite de passage* of entry into the discourse community, non-native English speakers (NNES) appear to be put at a significant linguistic disadvantage (Ammon, 2011; Englander, 2013; Hanauer and Englander, 2013; Hanauer, Sheridan, and Englander, 2019).

Markedly long grammatical subjects A related notable feature of scientific writing is the markedly long grammatical subjects (Gopen and Swan, 1990; Vande Koppel, 1994). Cognitive load is stretched by the lack of adherence to the principle of end-weight, in which long and complex elements are placed towards to the end of a clause to reduce the burden on short-term memory when reading (Biber, Johansson, et al., 1999, p.898).

Technical precision To avoid ambiguity precise, technical terminology is adopted. Subjective expressions, such as gradable adjectives tend to be avoided. An easy-to-understand example from the field of materials science is that stating the exact temperature of *39.8 degrees Celcius* is recommended rather than using gradable adjective *hot* (Parkhurst, 1990). Even simple words, such as *call, call back* and *get*, are assigned highly-specific technical meanings in information and computer science. This means that lay readers may assume they understand the meaning; but, in fact, in many cases may not as the lay interpretation is actually a misinterpretation (O'Donnell, 2013).

Table 2.13 lists some of the research conducted on research abstracts that discovered or aimed to investigate disciplinary variation in lexical realization in scientific research abstracts. As can be seen in Table 2.13, two thirds of the studies were conducted on corpora comprising fewer than 100 research abstracts. Regardless of the plausibility of the results, there is little empirical evidence to support the claims

TABLE 2.13: Studies investigating disciplinary variation in lexical realization of research abstracts

Year	Surname	Features studied or found	Discipline	Corpus size
1982	Rounds	hedging	BBS[a]	14
1985	Graetz	Past tense used to report research	Various	87
1987	Malcom	Present tense to refer to article cf. Graetz	unknown	unknown
1988	West	*that* nominal	Biology	unknown
1992	Salager-Meyer	Greater use of past time in verb phrases	Medical	84
2004	Lorés	Thematization	Ling[c] & AppLing[d]	36
2004	Hyland	Various, inc. hedging	Various	800
2005	Hyland & Tse	Findings & *that* clauses	Various	240
2007	Bonn & Swales	Linguistic variation between French and English	Ling[b]	60
2011	Tseng	Result & method use past tense	App Ling[b]	90
2011	Holtz	High lexical density, lack of modality	Ling[c], CompSci[f], Mech[i], Biol[j]	94
2013	Suntara & Usaha	Lexical features within moves	Ling[b] & AppLing[c]	200
2013	Tu & Wang	Prevalance of "be" verbs	Ling[b] & AppLing[c]	1000
2013	Ahmadi et al	More proficient writers used more MWEs[1]	Ling[b] & AppLing[c]	400
2013	Cao & Xiao	variation between Chinese & English writers	Various	10,741
2014	Esfandiari	Preference for present tense	CompSci[d]	32
2017	Jiang & Hyland	Function of metadiscursive nouns	Various	240
2019	Amnuai	Preference for present tense	Accounting	60

[a] Behavioural and brain science, [b] Linguistics, [c] Applied linguistics. [d] Computer science, [1] Multiword expressions

made. The conclusions from research studies with larger corpora (Cao and Xiao, 2013; Hyland, 2004; Tu and S. Wang, 2013), however, are based on firmer grounds, and give some interesting conclusions regarding the use of hedging, the verb *be* and variation in lexical realization between NNES and NES. However, none of the studies listed in this table focused on lexical realization within moves. This is most likely to do the time cost of annotating text for rhetorical moves prior to conducting lexical analysis. Given that no researchers have focused on the lexical realization within rhetorical moves across a range of scientific abstracts, a research gap that could be addressed was discovered.

2.6.3 Patterns of co-occurrence

> *"You shall know a word by the company it keeps!"*
> - John Rupert Firth

Firth (1957, p.11) famously stated and oft quoted, "[y]ou shall know a word by the company it keeps!" emphasizes the importance of co-occurrence of words.

Dalpanagioti (2018) shows how co-occurrence patterns move along a cline from narrow to broad. Figure 2.2 shows that the narrowest co-occurrence patterns are pure idioms, which can be classed as multiword expressions (MWEs). Collocation is the next level, which may also be realized with MWEs. Within the MWEs the idiomaticity may be pragmatic, semantic, syntactic or lexical (including grammatical). Colligation, semantic preference followed by semantic prosody are the broadest forms of co-occurrence.

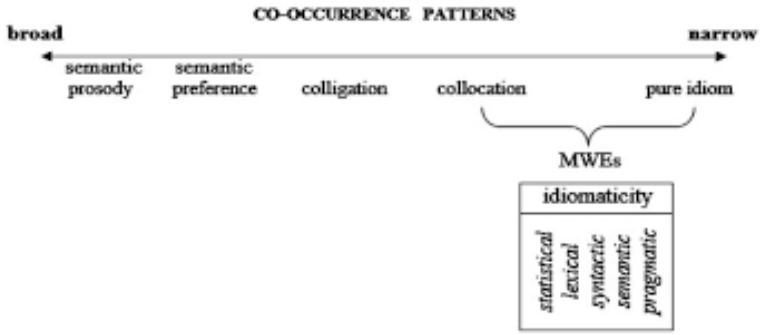


FIGURE 2.2: Cline of co-occurrence patterns approaches
Source: Figure 1 in Dalpanagioti (2018, p.427)

The term *collocation* may refer to a number of lexical features. Evert (2009) uses the terms *lexical collocations* and *empirical collocations* to clarify the differences between two types of associations. Lexical collocations are fixed expressions, which include idiomatic expressions and set phrases, and multiword units in which one or more units may vary while empirical collocations are used to refer to words that tend to co-occur.

Paice (1980, p.172) used the term *indicating phrases* to describe the common phrases or patterns of words using in research articles. Numerous linguists have written

about collocation (Barnbrook, O. Mason, and Krishnamurthy, 2013; Haswell, 1991; Hyland, 2008), colligation (Hoey, 2005; Gledhill, 2009) and multiword expressions (Biber, Conrad, and Cortes, 2004; Biber, Johansson, et al., 1999; Haan and Hout, 1986; Handford, 2007; Wray, 2005). Pawley and Syder (1983) claim these prefabricated phrases tend to be used in a limited number of permutations creating ordinary, idiomatic and natural sentences. Granger (2009, p.60) proposed the term phrasicon to describe this collection of phrases. The restricted set of phrases that are selected rather than relying on open choice (Sinclair, 1991, p.110) has been various defined, described and explained (e.g. principle of idiom in Sinclair (ibid., p.110) pattern Grammar in Hunston and G. Francis (1999); collostruction in Stefanowitsch and Gries (2003) and Gries and Stefanowitsch (2004).

The Academic Phrasebank (Morley, 2004; Morley, 2020) aims to provide a database of usable functional exponents for academic writers. This bank categorizes the functional exponents by function, such as comparing and contrasting, describing trends and explaining causality. There is a strong correlation between functions and moves, so the functions can be mapped to particular moves. This resource is, however, designed for general academic writing and so it is not clear how relevant the functional exponents are for scientific abstracts.

Collocational behaviours, also termed semantic prosody, can be discovered between words and the type of expressions that follow or precede them. Louw (1993, p.157) describes semantic prosody as "an aura of meaning with which a form is imbued by its collocates". Stubbs (2001, p.65) provides a more operationalizable definition of semantic prosody: "the relation [...] between a lemma or word form and a set of semantically related words". Sinclair (1991, p.112) discovered the verb *happen* is associated with "unpleasant things". Stubbs (1995) investigated this further and showed in a corpus of 120 million words that the lemma *cause* is associated with "unpleasant" collocations (e.g. trouble, accidents and death).

2.6.4 Keyness

Textual collocation has enabled the creation of genre-specific word lists, such as the first academic word list (AWL) (Coxhead, 2000, p.213). Hyland and Tse (2007) later discovered disciplinary variation in that AWL. Various other academic word lists have been proposed by analyzing the frequency of words in specific disciplines. However, none have analyzed the use of words within moves. Baker (2006a) notes that while simple word lists provide only information on frequency, key word lists provide a measure of saliency. Blake (2016) explains keyness as being:

> a measure of the frequency with which a word occurs in the corpus being analyzed (focus corpus) in comparison to another corpus (reference corpus). Words that occur more frequently show positive keyness and are commonly called key words (Scott, 1997). (pp.102–103)

Key words, simply put, are those words that are disproportionately frequent in the focus corpus when compared to a reference corpus. Key words can be identified statistically, counted and the differences between the key words can be quantified. Comparative occurrences in frequency per thousand or million words are usually used. It is worth noting that key word analysis focusses on lexical differences rather than lexical similarities (Baker, 2004). Kilgarriff (2012) shows that key word lists can be used as a practical and direct way to compare and contrast corpora or sub-corpora together. Scott (2009) asserts that any corpus can function as a reference corpus although different corpora result in different key word lists. Goh (2011) explains in detail how various generic factors may affect key word results. He concludes that key word analysis is robust, but differences in the genre and diachrony between the target text and reference corpus result in significant differences in the number of key words.

Blake (2016) also notes that the choice of concordancer, reference corpus and statistical test affect the key word lists generated. The choice of statistical test is problematic as a mathematical dilemma occurs when words appear in the focus corpus but not the reference corpus. This scenario results in the necessity to divide by zero in most of the commonly-used statistical formulae (Kilgarriff, 2012; Kilgarriff, 2009), which according to standard mathematical generates an undefined answer (Tsamir and Tirosh, 2002). Kilgarriff (2009), however, developed an alternative *Simple Maths* method for key words in which users select statistics according to the frequency range. The main difference is that either 1, 100 or 1000 is added to all counts per million for both the focus and reference corpora. This *trick* which is commonly used in mathematics for engineering solves the problem of potentially dividing by zero, but is not without issues. Church and Gale (1994, p.101) showed that the error rate caused by adding one for unseen bigrams may be as much as three orders of magnitude. This is a wicked problem in which there is no clear best answer, and so a compromise based on consideration of the evidence is the only feasible option.

Scott (2019) describes three kinds of key words, namely proper nouns, indicators of "aboutness" and high frequency words in the online manual for WordSmith tools. The indicators of "aboutness" are the lexical words that humans would recognise as key and related to the topic under discussion while the high frequency words are indicators of style rather than "aboutness", mainly due to their textual and grammatical functions.

There are a few small-scale studies focusing on key words occurring in one or two scientific disciplines. Suntara and Usaha (2013) provide a detailed analysis of language features within moves in the closely-related disciplines of linguistics and applied linguistics.

Hancioğlu, Neufeld, and Eldridge (2008, p.459) suggest that in addition to a word list what ESP practitioners need are "banks of lexico-structural items and collocates with genre-specific attributes and functions". Hyland (2008, p.20) echoes the calls of Nattinger and DeCarrico (1992), Lewis (1997), and Willis (2003) for more pedagogic

research into disciplinary specific research on lexical bundles. Although genre-specific word lists exist for different disciplines, at the time of writing there was no published record of genre-specific and move-specific multiword expressions for information theory, wireless computing, image processing and other highly technical disciplines.

2.6.5 Grammatical tense

> *Existence really is an imperfect tense that never becomes a present.*
> - Friedrich Nietzsche, On the Use and Abuse of History for Life, 1874.

Finiteness

Verb groups are used to describe actions, states or occurrences. Verbs may be *finite* and take a tense, or *non-finite* and not take a tense. Some verb types, such as *infinitives* and *-ing* forms may not carry tense. Consider the three sentences below.

(3) I wrote a program to solve this problem.

(4) I wrote a program, solving this problem.

(5) I wrote a program, which solved this problem.

There are two verbs, *wrote* and *solve* in each sentence. The first verb *wrote* carries a tense, specifically *past simple*, in all three cases (Examples 3, 4 and 5). The second verb *to solve* in Example 3 and solving in Example 4 do not carry a tense. Verbs that carry tense are called *finite verbs*. In Example 5, the second verb *solved* is finite as it takes a tense, *past simple*.

Tense

In a strict sense there are only two tenses, namely present and past. TimeML (Pustejovsky et al., 2003) automatically classifies verb groups into these two tenses. This classification, however, is insufficient to discriminate the usage of the twelve verb forms. Eight of these forms are tensed (past or present) while four are modalized (future). According to Biber, Johansson, et al. (1999), 85% of all verb forms in spoken English are tensed. The frequency distribution of tenses is skewed to favour simple forms that account for 90% of all cases. The four verb forms that are classed as *perfect progressive* make up less than 0.5% of all the verb forms (ibid.). Textbooks designed to help non-native speakers of English make extensive use of these twelve verb forms or grammatical tenses (Yule, 1998, p.54). The grammatical tenses combine tense, aspect and mood to convey 26 common grammatical meanings in context. These 26 categories can be more finely subdivided into 45 categories (Quirk and Greenbaum, 1993).

2.6.6 Twelve permutations

Grammatical tenses in this project are taken to be the twelve permutations of three time periods (present, past, future) and two aspects, namely progressive and perfect (ibid.) as shown in Table 2.14.

TABLE 2.14: Twelve grammatical tenses of the verb *do*

	Simple	**Progressive**[a]	**Perfect simple**	**Perfect progressive**
Present	Do	Be doing	(Has | have) done	(Has | have) been doing
Past	Did	(Was | were) doing	Had done	Had been doing
Future	Will do	Will be doing	Will have done	Will have been doing

[a] Continuous is used in some textbooks

Thus, the name of the tense contains at least two of three elements. The three elements are time period (*present, past, future*), perfectness (*perfect, non-perfect*) and continuity (*progressive, non-progressive*, i.e. *simple*). Table 2.14 shows a two-dimensional representation of tenses in a $3x4$ matrix, but a more efficient way would be through a three-dimensional $3x2x2$ matrix with tense, progressive aspect and perfect aspect as the dimensions.

The relative frequency of the tenses is skewed to simple tenses with 60% of all tenses in spoken English being *present simple*. Table 2.15 shows the results of a corpus study by Biber, Johansson, et al. (1999, pp.452-502).

TABLE 2.15: Frequency of the twelve grammatical tenses

	Simple	**Progressive**[a]	**Perfect simple**	**Perfect progressive** [b]
Present	60%	3%	3%	0%
Past	18%	1%	3%	0%
Modal [c]	12%	1%	3%	0%

[a] Continuous is also used in place of progressive
[b] Fewer than 0.5% of all verbs occur in perfect progressive
[c] Modal includes future forms using *will* and *going to*

The importance of tense to convey meaning accurately cannot be understated. Even non-linguists (Potter and Talukder, 2003) have published on the importance of tense in technical writing. The validity of a patent application may be questioned due to the ambiguity arising from the choice of grammatical tense. Tense is a central theme in coursebooks published for learners of English. The majority of published coursebooks designed for English language learning are structured so that learners are exposed to grammatical tenses in a systematic manner, frequently starting with present tenses. The choice of tense and verb impacts meaning, and so learners need to understand these grammatical tenses. There are two tenses in English: present and past, and there are three grammatical aspects, namely future, perfect and progressive. Coursebooks and pedagogic grammars tend to introduce twelve grammatical tenses (e.g. *present perfect simple, past progressive*, etc.). Grammatical

tenses are those pedagogic tenses that are used in textbooks to help non-native speakers of English understand which tenses to use in which situations.

Researchers have argued for decades about the use of past or present tense in scientific abstracts. Graetz (1985, p.125 cited in Swales, 1990, p.179) writes: "[t]he abstract is characterized by characterized by the use of past tense, third person, passive, and the non-use of negatives." Some of these claims are overgeneralized as counter examples are easy to discover.

Salager-Meyer (1992) found that although past tense was more frequently used, present tense was regularly used in medical abstracts to emphasize the generalizability of particular findings and foreground the new ground claimed. Malcom (1987 cited in Swales, 1990, p.179) found present tense is frequently used in abstracts to refer to the full article. More recently, some researchers investigate grammatical tenses even though the focus of their research differs. For example, Stotesbury (2003) investigated differences in evaluation among research abstracts in social sciences, humanities and natural sciences. Evaluation correlates to linguistic features, such as attitudinal lexis, modality and relevance markers (Hunston, 1994, p.198). Modality *inter alia* relates to the prophetic use of modal verbs to refer to the future aspect. The degree of disciplinary variation can be shown by considering the research of Samraj (2005), who noted variation in choice of tense and modality in the closely-related domains of conservation biology and wildlife biology. It is therefore plausible that greater variation would be discovered when comparing more disparate disciplines. Swales and Feak (2009) note that abstracts show considerable disciplinary and individual variation in the use of tense in the RESULT MOVE. Tu and S. Wang (2013) found in a corpus of 1000 research abstracts extracted from language-related journals (e.g. pragmatics and reading) that the tendency to use past or present tenses varied by journal. They also noted that prevalence for the use of copula *be*. Gledhill (2009) explains a contrastive function of the use of present and past tenses, claiming that in scientific writing present tense can be used to focus on given or accepted information while past tense is used for new information. Gledhill (ibid., p.5) also adds that perfect aspect can be used given in support of new information. Tseng (2011) in a study of 90 applied linguistics abstracts found that present tense was used in INTRODUCTION MOVES and DISCUSSION MOVE while METHOD MOVES and RESULT MOVE tended to use past tense. Biber and Gray (2013) note an increase in the frequency of past tense verbs in research articles in 2005.

There appears to be little consensus and much variation. With a more granular investigation of the twelve grammatical tenses within rhetorical moves across a broad range of scientific disciplines, it should be easier to get a clear view of the landscape of tense usage.

2.6.7 Section summary

There is a paucity of research specifically focusing on language features within moves in scientific research abstracts. What little research there is tends to focus on medical

and linguistic abstracts. There is little research into lexical realization within moves in research abstracts. Although the extraction and analysis of lexis and grammar is now much more easily achieved given the increase in power of fourth-generation corpus tools, the lack of availability of research abstracts annotated by move makes this an onerous task in terms of skills needed and time to be allocated. The cost of accurate annotation acts as an entry barrier.

Aside from Hyland (2004), there is no research that specifically investigates lexical realization within moves in a scientific research abstracts in over a range of disciplines. However, as his focus was not on scientific disciplines, only 400 scientific abstracts were included in his corpus.

There are currently no corpus studies that are able to provide insight into the relationship between moves and the lexis and grammar that are used to realize those moves in a research abstracts in a broad range of scientific disciplines, and none which deal with the hard-to-read technical disciplines within information science.

Although many research articles have been published which present key words lists for corpora, few are noteworthy. Nowadays, a simple corpus can be automatically created in less than an hour and at the click of a button key word lists can be calculated almost instantly. What is not available in the body of research is a key word list for rhetorical moves within a range of scientific disciplines. It could be that key words are more similar within disciplines than across rhetorical moves, or it could be that they are more within rhetorical moves than across disciplines.

There appears to be lack of agreement among researchers on the usage of tense in scientific research abstracts. The dimensions of disciplinary variation and rhetorical move need to be taken into account when creating a descriptive account of the usage of tense across a variety of scientific disciplines. Comparing two closely-related disciplines is interesting *per se*, but sheds little light on other scientific genres.

2.7 Chapter summary

Genre analysis and move analysis have contributed to our understanding of how communities of practice create and co-create meanings in texts through shared understandings, shared knowledge and a shared vocabulary. Despite its multifarious definitions and lack of a shared definition, genre is pervasive. Its complexity, however, belies the ease to which people feel that they understand what a genre is. Most people could confidently recognize prototypical examples of the genres of informal letter, formal report and academic essay. However, once genres become more technical, such as in the case of scientific research abstracts, lack of disciplinary knowledge and conventions complicates descriptive analysis.

To date there has been no large-scale multidisciplinary study on rhetorical organization focusing on the permutations of moves in scientific research abstracts. Some studies have identified the number of moves but none have focused on the patterns of moves *per se*. Previous studies have tended to be small-scale with corpora

of fewer than 100 abstracts or have used larger corpora but have provided little or no account of inter-annotator reliability. Disciplines within the humanities are far more accessible to lay readers than highly technical scientific disciplines. Readers with no knowledge of the subject area may be able to extract some relationships between phrases and clauses, but the exact meaning of the sentences is unlikely to be understood. This readability barrier no doubt explains the lack of attempts to deal with multiple different scientific disciplines in a single research study.

There is also a paucity of research on lexical realization within moves in scientific research abstracts. This no doubt stems from the substantial difficulties involved in annotating technical texts since without subject knowledge, annotator agreement falls substantially reducing the perceived validity of research results. Lexical realization within moves can only be analyzed once the crux of identifying rhetorical moves has been overcome. Given that the previous sections showed that there is no large corpus comprising different scientific disciplines of research abstracts annotated for rhetorical structure, no large-scale multidisciplinary study has ever investigated lexical realization at move level.

A number of studies have investigated keywords in abstracts in general and a few studies have compared two or three scientific disciplines together. The results are, however, not particularly revealing since it would be normal to expect medical abstracts to contain medical terminology and engineering abstracts contain engineering terminology. Since no-one has investigated the similarities and differences between the lexis and grammar in particular moves in different scientific disciplines, the research findings are *tabula rasa*. The research on the use of grammatical tenses in research abstracts is also rather patchy with coverage in only a limited number of scientific domains. To date there has been no attempt to map the usage of tense across rhetorical moves in wide range of scientific disciplines.

Although there are simple prescriptions, such as the four-move IMRD model, there is currently no framework or theory that provides a comprehensive picture of scientific research abstracts in terms of either the rhetorical organization or the lexical realization within the rhetorical moves.

In short, a descriptive corpus investigation into the lexical realization at move level in a range of disciplines could provide greater insight into the similarities among and differences between different scientific disciplines. From this, a framework could be developed that can help both researchers and practitioners to understand this high-stakes genre more clearly. This can become a valuable resource for both teachers of English for specific purposes and writers of research abstracts.

The review of the literature has confirmed that no study has reported in detail the patterns of rhetorical move organization in scientific research abstracts across a range of disciplines. In addition, there is no large-scale investigation of lexical realization within rhetorical moves in scientific disciplines. This study aims to fill this research niche by answering the main research questions:

1. What is the rhetorical organization of abstracts of research articles published in a broad range of top-tier scientific journals?

2. What are the lexical features of prototypical moves in abstracts of research articles in the selected scientific disciplines?

2.8 Research questions arising from literature

> *If I had an hour to solve a problem and my life depended on the solution, I would spend the first* 55 *minutes determining the proper question to ask for once I know the proper question, I could resolve the problem in less than five minutes*
> - (Attributed to) Albert Einstein

2.8.1 Preamble

The importance of the genre of scientific research abstracts written in English has been established as researchers, regardless of mother tongue, need to publish their research in top-tier English language journals. The concept of genre and the acceptance of the use of move as an ontological unit in the field of ESP has been established.

This section draws together the conclusions from the review of the literature, highlighting the gaps that are worthy of further investigation. Two main research questions emerged from the literature review. The first aim of this study is to investigate rhetorical organization which is the focus of the first main research question. The second aim relates to lexical realization within rhetorical moves, which the second main question addresses. The niche for each research question is introduced with reference to the gap in the literature that remains unfilled. Following an atomistic approach, each main research question is sub-divided into sub-questions that can be reformulated into testable hypotheses.

It has been shown there is a need for an in-depth exploration of scientific research abstracts in a range of scientific disciplines in order to help researchers, especially those with English as an Additional Language, disseminate their research more effectively and contribute to science using English as a *lingua franca*.

The main aim of this research is to describe and understand the rhetorical organization and lexical realization within moves in the genre of scientific research abstracts. Research abstracts of particular importance because of their functional role as portholes to the full research article. There is a plethora of genre studies on research abstracts in the humanities and in medicine, but the pure and applied scientific and engineering disciplines are largely under-researched. Once the rhetorical organization can be described accurately, a framework or schemata to help novice writers and teachers of scientific writing can be developed. This framework has a practical application in applied linguistics, namely to provide valuable information to help teachers and learners of research writing adhere to the generic conventions found in the descriptive analysis.

2.8.2 Research question 1

Although there have been a number of studies on rhetorical organization in research abstracts particularly in linguistics and medicine, few studies have focused on the multiple scientific domains and even fewer studies dealt with disciplines within information science. Given that most studies involved fewer than a hundred texts, it is highly unlikely that such a small sampling provides a representative sample of the whole population of texts, and so accusations of bias are easily cast. No large-scale annotated corpus study of rhetorical moves in research abstracts has ever covered a broad range of scientific disciplines.

The research on move structure is rather patchy with smatterings of small-scale studies providing glimpses but no overall picture of the state of scientific research abstracts. We know that textbook and self-help book writers advocate various forms of the IMRD move structure. We also know that abstracts that do not follow this pattern exist, but it is not clear how prevalent IMRD structure is in actual published research abstracts. Until now, we do not know what sequences of moves occur across a range of scientific disciplines. We still do not know if there are underlying trends or patterns that permeate different disciplines. These trends can only be discovered through statistical analysis of scientific texts that are annotation for moves.

If we could ascertain the patterns of moves that actual occur in published abstracts, this would provide valuable information that could help future cohorts of writers become familiar with the genre of scientific research abstracts in a more time-efficient manner.

The aim, therefore, is to create a framework or model that maps the commonalities and differences in similarity of patterns of move usage across a broad range of scientific disciplines. The main research question and its sub-questions are given below:

Main question

What is the rhetorical organization of abstracts of research articles published in a broad range of top-tier scientific journals?

Sub-questions

1. What moves occur in research abstracts in each discipline?
2. How frequent is each move in research abstracts in each discipline?
3. In what sequence do the moves occur in research abstracts in each discipline?
4. How frequent is each sequence in research abstracts in each discipline?
5. What are the similarities in rhetorical organization between the disciplines?
6. What are the differences in rhetorical organization between the disciplines?

2.8.3 Research question 2

There are very few studies addressing lexical realization within moves and no studies whatsoever that do so across a broad range of scientific disciplines. Taking the lack of large corpora of scientific research abstracts annotated for rhetorical organization into account, it is unsurprising that there is a paucity of research into lexical realization within rhetorical moves. Due to the fact that there has been little research on disciplinary variation at move level, the degree of differences or similarities between lexical realization with moves in different domains and disciplines is unclear. This study therefore aims to fill the gap in the literature by comparing and contrasting the lexical realization within moves between different scientific disciplines. Lexical realization will be investigated both lexically and grammatically following the assertion of Halliday (1992, p.64) that lexical questions get lexis-like answers and grammatical questions get grammar-like answers. To get lexical answers, key word analysis will be used while for grammatical answers, the pedagogic tenses will be used. key word analysis has been conducted previously to various extents but not across a wide range of scientific disciplines. There is a paucity of research on pedagogic tenses. This research also answers calls of researchers (e.g. Hyland, 2008; Willis, 2003) for more pedagogic research into disciplinary specific research on lexical bundles.

Answers to this main research question and its associated sub-questions should cast light on the degree of disciplinary variation within moves of scientific research abstracts. This data could be utilised by materials developers and teachers of research writing, particularly those with English as an Additional Language.

The main research question and its sub-questions are given below:

Main question

What are the lexical features of prototypical moves in abstracts of research articles in the selected scientific disciplines?

Sub-questions

1. Does the lexical realization differ between the same moves in different disciplines?
2. Does the lexical realization differ between different moves in the same discipline?
3. To what extent does the lexical realization differ between moves?
4. To what extent does the lexical realization differ between disciplines?

Chapter 3

Corpus linguistics

Corpus linguistics is one of the fastest-growing methodologies in contemporary linguistics.
- Stefan Thomas Gries

3.1 Chapter preview

The main thrust of this study is the application of corpus linguistics (Baker, 2006b; Biber, Conrad, and Rippen, 1998; Cheng, 2012; Evert, 2009; Hunston, 2002; Hyland, Huat, and Handford, 2012; Kennedy, 1998; O'Keeffe and M. McCarthy, 2010; McEnery and Wilson, 2001; McEnery and Hardie, 2012; Sampson and D. McCarthy, 2005; Sinclair, 1991; Sinclair, 2004c; Tognini-Bonelli, 2001) to address the gaps in the extant literature on the rhetorical organization and lexical realization of scientific research abstracts. The tools and techniques of corpus linguistics may be used to understand the frequency of individual moves and permutations of move sequences in the various scientific disciplines. A statistical approach can be used to identify similarities and differences in the frequencies of occurrence and frequencies of particular sequences of rhetorical moves between and among the disciplines. Once rhetorical moves have been identified, the frequency of phraseologies within moves across disciplines can be examined. This can be used to identify whether there are move-specific, discipline-specific, or genre-specific collocations or colligations.

This information can be used by materials writers and teachers of English for research writing to better address the needs of scientists. Should the content of the currently available pedagogic literature not accurately reflect the corpus findings, this may also reveal a niche in the academic publishing market.

This chapter is divided into five main sections, namely the rationale for adopting a corpus linguistics approach in Section 3.2, corpora in Section 3.3, corpus selection criteria in Section 3.4, corpus annotation in Section 3.5, and a chapter summary in Section 3.6.

In Section 3.2 begins by defining corpus linguistics. The benefits and drawbacks of adopting a corpus linguistic approach are discussed. The criticisms to such approach and their rebuttals are also presented. Section 3.3 defines what a corpus is and provides an overview of the different types of corpora. The four main corpus selection

criteria are discussed in Section 3.4. Each criterion is described and discussed in depth. The types of corpus annotation, their procedures and inter-annotator agreement are discussed in Section 3.5. Section 3.6 draws the chapter to a close.

3.2 Rationale for a corpus linguistics approach

3.2.1 Definition of corpus linguistics

> *The beginning of wisdom is the definition of terms.*
> - Socrates

Research approaches to writing can be broadly categorized into three types: reader-orientated, writer-orientated and text-orientated (Hyland, 2013, pp.192–194). The first two approaches prioritize human participants while the third approach prioritizes the text. This research adopts a text-orientated approach, specifically one that uses corpus linguistics.

Corpus linguistics is used to analyze corpus data statistically for linguistic features. English for Specific Purposes (ESP) is one of the areas in which corpus approaches have become mainstream (Belcher, 2006; L. Flowerdew, 2009). There is, however, not only one approach to corpus linguistics. Corpus linguistic approaches may be classified as corpus-driven, corpus-based or corpus-informed.

Corpus approaches in which the investigators begin with no prior assumptions or expectations are classed as corpus-driven. Tognini-Bonelli (2001, p.84-5) claims that the sole source of hypotheses should stem from the corpus itself. This corpus-driven approach is associated with neo-Firthians according to McEnery and Hardie (2012, p.6). The corpus-driven approach is somewhat akin to a grounded theory approach (Strauss and Corbin, 1997) using only the corpus as the source of data.

Corpus approaches that begin based on hunches, assumptions or expectations could be classed as corpus-based. Corpus-based approaches tend to use data from a corpus to "explore a theory or hypothesis" (McEnery and Hardie, 2012, p.6). In the corpus-based approach corpus linguistics is seen as a method. Figure 3.1 compares corpus-based and corpus-driven approaches. As can be seen, corpus-based approaches begin with a theory while corpus-driven approaches end with a theory.

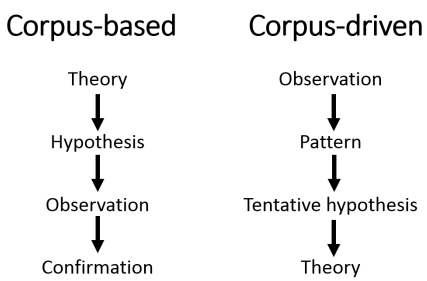


FIGURE 3.1: Corpus-based versus corpus-driven approaches
Source: Figure 8.6 in (.)187]cheng2012exploring

Corpus-informed approaches (ibid., p.17) tend to draw upon a corpus as a resource pool from which examples can be extracted. One common usage of corpus-informed approaches is in the development of materials for courses in English for Specific Purposes. A criticism of corpus-informed research is that examples may be viewed as being cherry-picked.

In addition to the three terms described above, three more terms are used that imply that a corpus linguistic approach is being used in conjunction with another approach. These joint approaches are called corpus-led, corpus-supported and corpus-assisted. Corpus-led approaches begin with data extracted from a corpus while corpus-supported and corpus-assisted approaches may not.

When selecting the approach, the appropriacy in relation to the purpose of the research remains the paramount factor. At the outset of this study, the researcher had some *a priori* expectations. The corpus was explored both using queries and during the annotation process, the whole corpus was read. As the researcher became more familiar with the content of the corpus, ideas for research questions and hypotheses arose, and so this study is situated firmly in the corpus-based approach of corpus linguistics.

3.2.2 Overview

This subsection introduces nine benefits of adopting a corpus linguistics approach, which are:

1. Extrospective description
2. Empirical approach
3. Scope of scrutiny
4. Objective analysis
5. Adaptability
6. Non-linear reading path
7. Pattern discovery
8. Technological accessibility
9. Modifiability

After this, the general trend to using corpus linguistics in many linguistic fields is discussed. Four criticisms of corpus linguistics are raised and responses provided. The subsection concludes by asserting that the benefits outweigh any drawbacks, supporting the case for using corpus linguistics in this research.

3.2.3 Benefits of a corpus linguistics approach

This subsection describes the benefits and drawbacks of adopting a corpus linguistic approach. In short, the benefits far outweigh the drawbacks, some of which are now obsolete given advances in technology while others can be ameliorated. The benefits listed are not presented in any hierarchy nor as being mutually exclusive as each benefit interacts with other benefits. The overarching benefit of adopting a corpus approach is the potential for novel insights based on statistical analysis of a large dataset.

Benefit 1: Extrospective description

A few decades ago, linguists tended to rely on intuition and introspection to reflect on language usage (Sinclair, 1991, p.1; Scott and Tribble, 2015, p.3). The underlying assumption was that it was possible for individuals to access their own cognitive system (Schütze and Sprouse, 2013, p.29). Fillmore (1992, p.35) used the term "arm-chair linguist" to describe those who dream up examples and subsequently present those examples as facts. The reflection naturally began with their own linguistic repertoire and, at times, was extended to the linguistic repertoires of other members of their speech community. The move from introspection to extrospection more easily enables researchers to adopt a descriptive approach based on data rather than introspective intuitive guessing (Holtz, 2011, p.36). Corpora contain language used for a communicative purpose rather than the language that prescriptive grammarians may accept as correct. This brings the focus on actual description of instances and clusters and not idealized language, i.e. the focus is on linguistic performance, not linguistic competence (Leech, 1992, pp.107–111) or on *parole*, not *langue* (Saussure, 1966).

Benefit 2: Empirical approach

The methodology of corpus linguistics is empirical (Leech, 1992) and measurable. The corpora used are not figments of imagination or pure theory, but are artifacts (texts) that existed prior to the research. The texts contain features such as words and phrases that can be counted, and so provide researchers with data points. Two cornerstones of the scientific method, namely replicability and falsifiability, are eminently achievable using corpus methods. An added benefit is that the artifacts of the research (e.g. corpora), the methods of added value (e.g. annotations) and methods of analysis (e.g. statistical code) can be shared with other researchers, adding transparency to the research methodology.

Benefit 3: Scope of scrutiny

Hunston (2013) describes depth as one of the flavours of corpus linguistics. By this she means that a corpus approach allows researchers to scrutinize instances or clusters of

instances of language at various levels of granularity. Gries (2011, pp.188–189) also points out that researchers can move along a cline of granularity from coarse-grained to fine-grained, contextualizing features of interest. This ability to zoom in and out of particular language features, or nodes, and see the co-text helps contextualize any analysis of features. Given the huge volume of processable data (termed "extent" by Hunston, 2013), it is useful to have the ability to focus on particular nodes and pan out to their co-text, moving between serialistic and holistic, or microscopic and macroscopic viewpoints.

Benefit 4: Objective analysis

To a certain extent, objectivity can be achieved in terms of "trusting the text" (Sinclair, 2004c) and using the corpus as evidence. As Peshkin (1988, p.20) notes, escaping subjectivity is perhaps unrealistic, but the use of the corpus goes someway towards "taming one's subjectivity". It should be acknowledged that the creation of the corpus and the choices of investigation avenues to pursue no doubt contain both known and unknown biases. The frequency focus of corpus linguistics, namely counting language features enables quantification (Gries, 2011, p.184). This quantification transforms a text or collection of texts into statistical data. That dataset can then be visualized using tables, graphs and charts to discover or investigate patterns. Statistical analysis can be used to discriminate marginal from "central and typical" usage of language (Gries, 2006, p.191; Hunston, 2002, p.42; Sinclair, 1991, p.17) and test reliability (Gries, 2006, p.191). The patterns discovered are representative of the dataset, therefore avoiding accusations of cherry-picking. Prototypical patterns of usage discovered, e.g. rhetorical structures, can then be used as pedagogic models in the language classroom (T. Upton and M. Cohen, 2009, p.22).

Benefit 5: Adaptability

A corpus linguistic approach can be used with both quantitative and qualitative models of language (Leech, 1992). One way is to complement the quantitative findings arising from corpus analysis with qualitative findings (L. Flowerdew, 2004; O'Keeffe, 2007). Corpus linguistics can be applicable to almost any theoretical framework (Thompson and Hunston, 2006, p.2), and has shed light on descriptive linguistic features in all branches of linguistics (Leech, 1997, p.9; Biber, Conrad, and Rippen, 1998, p.11).

Benefit 6: Non-linear reading path

Hunston (2002, p.3) notes that although a corpus contains no new information about language, textual analysis software enable new insights by viewing language from a different perspective. One difference between typical text linguistics and corpus linguistics is the method of *reading* a text. In text linguistics researchers *read* one text at a time using extensive and/or intensive reading, reading for gist and reading for

specific details. However, in Corpus linguistics researchers can *read* multiple texts simultaneously usually in a non-linear manner. Tognini-Bonelli (2010, p.19) provides a clear tabular summary of key differences as shown in 3.1. Kress (2003) discusses non-linear reading, focusing on multimodal digital hyperlinked texts. However, extracting information out of a text using corpus tools can also be considered a form of non-linear reading. Adopting a corpus linguistic methodology enables researchers to access the text via non-linear reading paths; thus, enabling researchers to extract data that traditional linear reading may not have revealed.

TABLE 3.1: A qualitative comparison of a text versus a corpus

A text	**A corpus**
Read whole	Read vertically
Read horizonally	Read for formal patterning
Read for content	Read for repeated events
Read as a unique event	Read as a sample of social practice
Read as an individual act of will	Gives insights into *langue*
Instance of *parole*	Not a coherent communicative event
Coherent communicative event	

Source: Tognini-Bonelli (2001, p.3) and Tognini-Bonelli (2010, p.19)

Benefit 7: Pattern discovery

A key strength of corpus linguistics is its focus on "probabilities, trends, patterns, co-occurrences of elements, features or groupings of features" (Teubert and Krishnamurthy, 2007, p.6). Statistical data can be visualized for example using boxplots with notches in the programming language R. This visualised data can revel hitherto undiscovered patterns in the dataset. Researchers are, therefore, able to make novel insights by investigating aspects of language that introspection or intuition alone would be unlikely to identify, such as aspects of collocation, frequency, semantic prosody and phraseology (Hunston, 2002; Louw, 1993; McEnery and Hardie, 2012; Reppen, 2012; Stubbs, 2001).

Benefit 8: Technological accessibility

With the reduction in the cost of processing power, the ease of access to the internet and the proliferation of online corpus tools, techniques that were once only available to computer programmers have become accessible to linguists with little or no programming expertise. The most popular corpus tool is a concordancer. McEnery and Hardie (2012) classified concordancers into four generations although the first two generations are now obsolete. An overview of third and fourth generation concordancers is provided in 3.2. Third-generation concordancers usually need to be installed or executed after downloading. Popular concordancers include AntConc (Anthony, 2019) and Wordsmith Tools (Scott, 2019). Fourth-generation concordancers, such as Sketch Engine (Kilgarriff et al., 2014) are typically online, deal with large corpora and are far more powerful than third-generation concordancers.

TABLE 3.2: Current generations of concordancers

	Third generation	Fouth generation
Location	Personal computers	Web servers
Size of corpora	Small corpora (low millions)	(100 million upwards)
Examples	AntConc, UAM Corpus Tool, Wordsmith Tools	CQPweb, Sketch Engine, W-matrix

Source: Table 1: Current generations of Concordancers in Blake (2016, p.103)

As frequency and recurrence are of importance in language (Stubbs, 2007), corpus linguistics can be harnessed to analyze evidence through statistical analysis. Corpus linguists approach texts from this perspective by compiling corpora and counting frequencies of particular language features. Sinclair (2004b) points out the centrality of lexical items in corpus linguistics. This is because most corpus tools or concordancers are designed to find patterns at word level, e.g. frequency lists, key word in context, etc.

Benefit 9: Modifiability

To quote Pillai (2017, n.p.), "Modifiability is the degree of ease at which changes can be made to a system, and the flexibility with which the system adapts to such changes". Although many out-of-the-box corpus tools exist, creating tailor-made programs in Python or R can finetune the analysis, meeting the exact needs of a corpus linguist. For this reason, the R programming environment is gaining in popularity. R is a software environment for statistical computing and graphics which is far more powerful than any corpus tool. The trend to more rigorous use of statistics is also most likely linked to the increase in use of programming languages such as Python and R. The growing popularity of R among corpus and computational linguists has also changed the technoscape (Appadurai, 1990) of corpus linguistics. Previously, many corpus linguists used concordancers created by a handful of software developers. Nowadays, many corpus linguists are adapting or creating their own programs in Python and R.

Trend

Although not a benefit, the trend in many areas of linguistic publishing has shifted from intuition and introspection to evidence-based analysis. Anthony (2016) showed an increase in corpus approaches in the field of ESP by analyzing the journal article targets for the English for Specific Purposes journal. Trends are an indicator of popularity. Popularity *per se* is not indicative of any truth, but it is an indicator that researchers are able to discover publishable quality results using corpus approaches. This trend is a corollary of the increase in internet connectivity, computer-savvy researchers and the development of easy-to-use corpus tools. Previously, creating a small corpus was an onerous task, but nowadays a simple web-crawled corpus can

be created in a few clicks using corpus tools, such as Sketch Engine (Kilgarriff et al., 2014).

The move to evidence-based research has been enthusiastically received in many sub-domains of linguistics. According to Gries (2015), corpus linguistics is one of the fastest-growing methodologies in linguistics. Gries (ibid.) also notes the increase in the statistical nature of corpus linguistics in this millennium, citing the frequency of research reporting monofactorial statistical tests and a rise is those reporting multifactorial statistical tests. There is an increase in the number of corpus studies in many fields of linguistics, presumably as corpus approaches have been able to reveal novel insights.

3.2.4 Criticisms of corpus linguistics approach

Criticisms of adopting a corpus linguistic approach, however, can be categorized into four broad aspects. To quote (Handford, 2012b, p.255), the four aspects are:

- Corpus data are decontextualised data (Widdowson 1998, 2000)
- Corpora require a bottom-up approach (Swales 2002)
- Corpora are quantitative, number-crunching tools (see Baker 2006, p.8) [....]
- The bigger the corpora the better (Sinclair, 1991; Stubbs, 1996)

Each of these issues is discussed in turn in the following paragraphs.

Issue 1: Decontextualized data

Simply put, the traditional assumption is that discourse analysts focus on the language data in context while corpus linguists focus on the quantifiable data out of context. However, as Virtanen (2009, p.1062) suggests that: "[c]ombining methods from corpus linguistics and discourse analysis [...] should smooth the way for a happy relationship between the two areas of study".

In the early days of corpus linguistics, it was difficult to access details of the wider context of language features, particularly so for general corpora. However, with technological improvements co-text and context can be accessed instantly with many user-interfaces. Combined approaches, as recommended by Biber, Connor, and T. A. Upton (2007), such as using top-down discourse analysis to classify rhetorical moves followed by statistical analysis of the language features discovered in each category, do not suffer from the problem of decontextualization. Another way to ameliorate the decontextualization issue is for investigators to familiarize themselves with language features of interest in the dataset to be analysed. The aim of this familiarization is to "ground quantification [...] of linguistic selections and constructions, through qualitative analysis of discourse" (Drew, 2004, p.221), using examples of the linguistic features found in a corpus (McEnery and Wilson, 2001, p.71). Widdowson's argument

relates to general corpora in which the analyst is not aware of the context of specific language usage other than being able to identify a a few words to either side of any particular features when using Key Words In Context. However, in the case of specialized small corpora, analysts may be familiar with whole texts, possess subject-specific knowledge and be able to draw on exogenic knowledge during analysis. Schiffrin (1994, p.419) asserts the need to understand language in context, or as she puts it, "To understand the language of discourse...we need to understand the world in which it resides".

The quantitative-qualitative research debate exists in many different scientific domains. In linguistics, corpus linguists start their research from a quantitative standpoint while many sociolinguists and applied linguists may start from a qualitative perspective. Both paradigms have advantages, which explains the rise in mixed-methods research. As Hyland (2012a, p.30) puts it, "[c]orpus analysts are, however, sensitive to criticism that they treat texts as artifacts abstracted from contexts. As a result, they have increasingly added a focus on *action* to balance the focus on *language*." Interviews with authors, readers and specialist informants are used to contextualize or elaborate the meanings. The use of specialist informants to provide expert advice on texts in specialized technology genres is commonplace (Hyland, 2012b; Lieungnapar and Todd, 2011).

Issue 2: Bottom-up approach

L. Flowerdew (2009, p.395) notes that the examination of concordance lines using key word in context (KWIC) forces researchers to analyze discrete language features in a rather atomistic way. However, this is easily ameliorated by accessing a larger segment of text or the whole document. Kaltenböck and Mehlmauer-Larcher (2005, p.71) point out that concordancers cannot access certain textual features, such as the macro-structure of a text. However, by judicious use of top-down annotation, the moves and sub-moves that combine together to form such structures can be identified and analyzed.

Issue 3: Frequency focus

Language is complex and so simply counting features does not provide the whole picture. It could be that one occurrence of a single word may be far more salient than multiple occurrences of a different word. Salience cannot be accounted for by counting. It is therefore necessary to understand the context and co-text of features that are being counted. Handford (2012a, p.17) notes the importance of understanding the link between text and context when working with professional discourse. Another issue with the quantitative focus of corpus linguistics is the method of statistical analysis selected. Gries (2006) notes various statistical issues regarding quantitative analysis undertaken by corpus linguists. Some statistical issues include the difficulty to clearly define the population to which the sample should be generalized to.

Issue 4: Bigger is better

Gries (2006) notes the concerns regarding the way that the data is accumulated and organized. This is because corpus specification decisions may negatively impact the representativeness of a corpus. Corpus size and representativeness are discussed in depth in 3.4. It is worth noting that generalizations, such as *bigger is better* are like stereotypes: they hold true in many cases, but not all cases. To provide a counter example showing that *bigger is not always better*, a small specialist corpus targetting a niche genre can provide more novel insights into the specific language features of that genre than any large corpus could (Koester, 2012, p.69; O'Keeffe, M. McCarthy, and Carter, 2007, p.198). Gries and Newman (2013, p.259) provides multiple examples of small specialist corpora that were created for specific research goals, such as one used to investigate the phonetic reduction of *that*.

3.2.5 Section summary

This discussion of the advantages and disadvantages has raised some strong arguments for the adoption of a corpus linguistic approach to this study. In short, corpus linguistics is a methodology that enables empirical quantitative research to be conducted on texts using descriptive or analytical statistics. There are numerous benefits to using a corpus linguistics approach while the criticisms of such an approach can be ameliorated by taking into account by compensating for the potential drawbacks.

3.3 Corpora

3.3.1 Corpora Overview

This section begins by defining the term corpus (Subsection 3.3.2) and providing an outline of the various types of corpora (Subsection 3.3.3). The following subsection (Subsection 3.3.4) discusses specialist corpora, which is particularly important in this study. The summary (Subsection 3.3.5) highlights how a corpus can serve as the evidence source for objective statistical analysis (e.g. measurable and replicable), and lists some corpus selection criteria, such as size, representativeness and balance. The corpus selection criteria are discussed in depth in Section 3.4.

3.3.2 Definition of corpus

A corpus can be defined as "a collection of naturally-occurring language text, chosen to characterize a state of variety of a language" (Sinclair, 1991, p.171). McEnery and Wilson (2001, p.31) refine this definition by amending the concept of being *machine-readable* (i.e. plain text) so that the corpus can be read by concordancing software. A common usage of corpora is to access prototypical usages of language (Hunston, 2002, p.42; Sinclair, 1991, p.17). Hunston (2002, p.243) adds a new layer of meaning by stating that "a corpus is defined not by what it contains but by how it is used".

3.3.3 Types of corpora

The taxonomy of types of corpora is extensive and includes monitor or reference corpora (e.g. Corpus of Contemporary American English), parallel translation corpora (e.g. Japanese-English) and comparable corpora (e.g. different languages or varieties of languages). Balanced/representative corpora, such as the British National Corpus (BNC), contain texts collected based on pre-defined specifications with the aim of replicating a language or language variety. Diachronic corpora contain texts that were produced in different or consecutive periods, such as the Time magazine corpus housed in Brigham Young University collection (Davies, 2007). A specialist corpus targets a narrower focus, such as a particular genre, text type, discipline or combination thereof.

Corpora may be named by how they are created. Transcribed corpora are used when representing spoken language in a written form, (Optical character recognized) OCRed corpora can be used for handwritten texts and non-digital documents. Digital texts, such as websites, blogs and online journal, can be downloaded manually or automatically downloaded using specialist software or scripts. Automatically downloaded corpora are known as (web-)crawled corpora. Lee (2010, pp.109–116) provides an extensive list of the major corpora for English language arranged by categories, such as historical, multimedia and parsed.

3.3.4 Specialized corpora

Whether a general or a specialized corpus is selected depends primarily on the purpose of the research. A corpus needs to be representative of language under investigation (Reppen, 2012). Purpose is also the critical factor when designing the size of a corpus (M. Nelson, 2010, p.54). General corpora, such as the Corpus of Contemporary American English and the British National Corpus are a valuable resource for investigations into, for example, common collocations of verbs related to emotion, but they are not useful resources to investigate the use of verb phrases in scientific research abstracts. Investigations into a particular genre, text type, time or period are more likely to necessitate a specialist corpus, which may need to be specially constructed if no suitable corpus exists. "Specialized corpora, which are generally less than a million words and sometimes much smaller, tend to benefit from contextual information to allow for interpretation of the genres in question through a combination of quantitative and qualitative approaches" (Handford, 2010 cited in Handford, 2012a, p.15). Numerous specialist corpora have been compiled to answer research questions that more general corpora were unable to address. Lee and Swales (2006) lists a number of specialist academic and professional corpora, such as the Michigan Corpus of Academic Spoken English, British Academic Spoken English corpus, Reading Academic Text corpus and the Wolverhampton Business English Corpus.

Corpus construction necessarily needs to be guided or driven by the specific goals of the investigator. A specialized corpus as one that comprises texts of a specific type, such as scientific research abstracts. Linking language patterns to a particular context through the usage of a small specialized corpus "brings into clear focus signature uses of language" (O'Keeffe, M. McCarthy, and Carter, 2007, p.182). Koester (2012, p.67) echoes this and notes that:

> "smaller, more specialized corpora have a distinct advantage, allowing a much closer link between the corpus and the contexts in which the texts in the corpus were produced...[and] give insights into patterns of language in particular settings".

L. Flowerdew (2004, p.19) states that small corpora contain fewer than 250,000 words. No rationale is given for this cut-off point and, as such, seems rather arbitrary. Given that small specialized corpora may represent a particular genre more accurately than much larger general corpora, the patterns of genre-specific lexical and grammatical structures are likely to be easier to ascertain with a tailor-made small specialized corpus (Bowker and Pearson, 2002; Koester, 2012, p.69; O'Keeffe, M. McCarthy, and Carter, 2007, p.198). Kennedy (1998, p.68) points out that when compiling a corpus, the corpus builder must keep a firm eye on the quality as well as quantity of data.

3.3.5 Section summary

In short, empirical evidence for a statistical study of language can be obtained from a corpus. The selection of the type of corpus is determined by the purpose of the research. In this study, a specialist corpus containing only research abstracts from the target publications is most likely to provide better quality data than a larger more general corpus.

At the outset of this study, to the best of my knowledge, there was no freely-available balanced representative corpus of research abstracts that covered a wide range of scientific disciplines, and so a corpus needed to be constructed. Given that research abstracts are freely available online, a digital or a crawled corpus can be created. The purpose of the research needs to be considered first when determining the size and balance of a representative corpus. The corpus selection criteria are discussed in depth in the following section.

3.4 Corpus selection criteria

3.4.1 Overview

This section first introduces the problems facing corpus designers and provides a glimpse into the decisions that need to be made. Four core criteria (3.4.2) for selection are discussed. Four subsections then consider the four criteria.

- Subsection 3.4.3 – Size
- Subsection 3.4.4 – Representativeness
- Subsection 3.4.5 – Balance
- Subsection 3.4.6 – Sampling frames, including samples, population and units

The summary in Subsection 3.4.7 draws out the key concepts that need to be considered in design of a specialist corpus suitable for the purpose of this study.

3.4.2 Four core criteria

Corpus design is a vexed issue. Design criteria are frequently critiqued and personal biases of corpus designers can easily skew any corpus, and in turn skew any results derived. Terms such as size, sampling, representativeness, balance are frequently used to justify corpus design. Many corpus compilers simply cite precedents as the basis of authority for their design, provide scant descriptions of the population and insufficient data for a statistician to analyze. As Sinclair (1991, p.13) noted "the results are only as good as the corpus". Among the many prescribed guidelines regarding corpus design, three intertwined aspects that are frequently mentioned are size, representativeness and balance. The key issue of representativeness is inextricably intertwined with sampling, size and balance.

Fillmore (1992, p.35) asserts that despite the deficiencies in all corpora regardless of size, without corpora many facts about the nature of language would remain undiscovered. A case in point is semantic prosody or discourse prosody which can be either positive or negative depending on whether a word tends to collocate with words with positive or negative connotations (Stubbs, 2001).

As stated in the previous section, purpose is the paramount consideration in the selection or creation of a corpus (M. Nelson, 2010, p.54). Corpus developers must consider a number of interrelated factors when designing a corpus. Corpus designers understand the intricacies of the corpora they create. This is in contrast to readers of research manuscripts based on the corpora who tend to remember only the size and content of corpus, presumably as these aspects are more tangible. This may be a remnant of one of the early claims regarding size, namely that biggest was considered the best (Kennedy, 1998).

The content criterion is related to the subject matter of the texts. Sampling is concerned with the selection of the sampling unit and sampling frame. Representativeness contains two aspects:

1. whether the sample (corpus) is generalizable to the population, and
2. whether the full range of target features (e.g. rhetorical moves in this study) occurs in the sample (corpus).

Balance is used to describe whether the balance of texts within the sample (corpus) accurately reflects the balance of texts with the population. Balance is assessed in relation to the purpose of the study and could be described in terms of text types of variables. Researchers classify representativeness, balance and sampling differently and so some criteria are at times subsumed within another criterion.

3.4.3 Criterion 1: Size

Size and sampling combine together to affect representativeness, but a finite corpus regardless of size is not able to exemplify all the language patterns in their normal proportions (Sinclair, 2008, p.30). The size of a corpus is dependent upon the research focus and practical considerations, such as the availability of time, money and resources (Dörnyei, 2007, pp.309–310). Purpose is the critical factor when designing the size of a corpus (M. Nelson, 2010, p.54). Corpus size is a vexed issue and the notion that one size will suit all purposes is a myth (Carter and M. McCarthy, 2001). Early claims regarding size distill down to the bigger, the better (Kennedy, 1998).

For example, Sinclair (1991, p.18) advocated that corpora should be as large as possible due to the large percentage of *hapax legomena*, i.e. words that only occur once within a context. *Hapax legomenon* (Coulthard and A. Johnson, 2007) is a word that only occur once within a particular context or corpus. Scott and Tribble (2015, p.26) notes that *hapax legomenon* occurring in the British National Corpus tend to be proper nouns, but some are foreign words, alternative word forms and typos. Baroni and Ueyama (2006, p.1) point out that:

> Because of Zipfian properties of language, even a large corpus such as the BNC [British National Corpus] contains a sizeable number of examples only for a relatively limited number of frequent words, with most words of English occurring once or not occurring at all.

A decade later Sinclair (2001) had changed his stance and noted that small corpus studies are not necessarily intrinsically bad. In defense of a small corpus, Biber (1990) showed that reliable results can be obtained from a corpus containing just a thousand words. Domain-specific research is more likely to be able to harness smaller corpora (Hunston, 2002, p.15). This view is supported by Tribble (1997, p.5) who states:

> if one wishes to investigate the lexis of a particular content domain (e.g. health) a specialist micro-corpus can more often be useful than a much larger general corpus. For example, in the written component of the BNC Sampler (1,000,000 words) there are no instances of 'cancers'. An Encarta® micro-corpus of health articles (24,805 words) gives 33 usefully contextualised examples!

Linguists working on smaller, more specialized corpora argued that the smaller balanced corpora can yield more fruitful results for particular pedagogical purposes (J. Flowerdew, 1996; L. Flowerdew, 1998). Koester (2012, p.67) notes that:

> smaller, more specialized corpora have a distinct advantage: they allow a much closer link between the corpus and the contexts in which the texts in the corpus were produced... [and] give insights into patterns of language in particular settings.

According to Lee (2008), also cited in Handford (2012b, p.258)), "it seems plausible that the more specialised the genre, the smaller the corpus can be". As scientific research abstracts are extremely specialized, harnessing distinct lexical sets and rhetorical patterns, it is likely that a comparatively small corpus could yield representative results.

Size is, however, a subjective construct. The perception of large, particularly in electronic terms is ephemeral (Sinclair, 2001, p.ix). Although finite size appears relatively easy to measure, the issue is what exactly should be measured when evaluating size, since there are so many possible units which can be quantified, such as: number of texts, sentences, words, tokens, syllables, letters, or even bytes. The reference corpora of the past now appear minuscule in comparison to reference corpora compiled more recently with the Brown corpus standing at one million words while the 2013 Google book corpus is 155,000 million words. (see Table 3.3).

TABLE 3.3: Size of reference corpora

Date	Name of corpus	Total number of words (million)
1960	Brown Corpus	1
1970s	Lancaster-Oslo-Bergen Corpus	1
1994	British National Corpus	100
2012	Corpus of Contemporary American English	450
2013	Global Web-based English Corpus	1800
2013	EnTenTen Corpus family	10000
2013	Google Book Corpus	155000
2018	NOW (news on the web) Corpus	5500

Biber (1988) states that degree of variation within a genre should determine the size of the sample from within that genre. Sample or corpus size depends on the linguistic features (M. Nelson, 2010, p.98), to be investigated. This means that without creating an initial corpus and investing its linguistic features, it is not possible to know whether the corpus size is appropriate. Biber (1993), also cited in Kennedy (1998, p.69), states "empirical research on the pilot corpus should be used to confirm or modify the design parameters". Corpus design can therefore be seen as an iterative process in which a corpus is created and then its suitability evaluated. If the corpus needs to be altered in terms of size, balance or any other factor, this can be undertaken; and then, the corpus re-evaluated. This process continues until either the limitations of time and finance are reached, or appropriate size, balanced, etc. has been achieved.

3.4.4 Criterion 2: Representativeness

The goal of the corpus designer is to create a corpus that is "maximally representative of the language variety under consideration" (McEnery and Wilson, 2001, p.24). A key consideration is that the corpus is representative (Biber, 1993; Koester, 2012; Reppen, 2012). Corpora are judged on the merits of their representivity (O'Keeffe, M. McCarthy, and Carter, 2007); when researchers generalise from results obtained from a corpus that is not representative, their results are likely to be met with skepticism. Otherwise, there is likely to be less skepticism.

Váradi (2001) highlights the failure of corpus linguists to adequately define balanced and representative corpora. Leech (2007, pp.143–144), however, asserts that despite their ill-defined status, problematic nature and lack of attainability, the concepts of balance and representative are key considerations. Advocating a "gradual approximation" to the goal, he notes that there is a "scale of representativity, of balancedness, of comparability". In a similar vein, McEnery and Hardie (2012, p.10) explain that "[b]alance, representativeness and comparability are ideals which corpus builders strive for, but rarely, if ever, attain. In truth, the measures of balance and representativeness are matters of degree." Representativeness remains a noble goal, but is in fact a nebulous notion (Hunston, 2002, pp.28-30). Kennedy (1998, p.62) notes the difficulty in selecting texts to create a corpus that is truly representative of all the possible genres, fields or topics. As such, compromise between the ideal corpus and the actual corpus is inevitable (M. Nelson, 2010, p.60). Leech (1991, p.27) asserts that a corpus is representative when findings based on its analysis are generalizable to a "larger hypothetical corpus". Given the impossibility of collecting all the available texts, it is necessary to select a "maximally representative" sample of the texts to minimize bias (McEnery and Wilson, 2001, p.32). Gries and Newman (2013, p.257) note the need to avoid creating a corpus that confirms "the analyst's pre-existing expectations".

Claims of corpus linguists may be rejected on the basis that their corpora are not representative of their intended genre or register. Yet, as Tognini-Bonelli (2001, p.57) notes, the evaluation of representativeness is itself a subjective judgement. Natural scientists use the scientific method and deductive reasoning to infer sound conclusions using valid deductive arguments based on premises that are unable to be falsified. However, many social scientists, including corpus linguists, tend to use hypothesis testing on what statisticians would class as relatively small samples and infer conclusions using inductive reasoning to an ill-defined population. In the early days arguments against the use of corpus linguistics such as "[a]ny natural corpus will be skewed" (Chomsky, 2002, p.159), were commonplace. However, with the acceptance of cogent arguments for the use of corpora and the concomitant increase in the sophistication of corpus tools, the use of corpus linguistics is now used in many linguistic disciplines. The unprecedented breakthroughs in identifying new and novel concepts have brought about a sea change in many linguistic disciplines.

Representativeness in this study is defined as having two criteria, namely generalizability and range. For a corpus to be generalizable, the texts in the corpus must accurately represent those in the population to which the generalization is applied. In order to avoid accusations from statisticians, population can be defined in a measurable way. Thus, there is the actual population, which may be extremely difficult to estimate with any degree of accuracy (e.g. all journal articles in a particular discipline) and the measurable population (e.g. all journal articles published in one journal for a decade), which can be estimated with more certainty. For a corpus to be considered as fulfilling the range criterion, the full range of variation in the target language features should be present in the sample. Biber (1993) advocated evaluating representativeness on range of distribution of (1) text types and (2) linguistic features of interest. He claims that if text types are not representative, then linguistic distributions will not be representative. Biber (ibid.) further asserts that linguistic representativeness is dependent on three variables: situational representativeness, the number of words per sample and the number of samples per corpus. Representativity can be considered as being internal, i.e. within the sample (corpus); and external, i.e. related to the population. Biber, Conrad, and Rippen (1998, p.250) declared that representativeness can be investigated empirically. This can be achieved by providing statistical evidence to shore up accusations related to the generalizability and range of the corpus and its internal and external consistency.

A corpus is representative if the variation of the features of interest is similar within the corpus itself and among the population. A corpus, C, is in fact a sample, S, of a population, P. Populations whose parameters are known can be called measurable, but it may be the case that the actual total population to which generalizations will be made are not measurable. Leech (1991, p.27) asserts that a corpus is representative when findings base on its analysis are generalizable to a "larger hypothetical corpus", i.e. a measurable population. There is external consistency when the measurable population is representative of the actual population. However, this assumption which cannot be proven can be declared as a limitation rather than hidden as an unstated assumption. The second part of external consistency is that the sample corpus is representative of population in terms of the full range of target features. Given the inductive nature of this reasoning, the outcome can never be proved certain, but it can be disproved. The degree of representativeness can be estimated by comparing the means and standard deviations across the corpus. If the means and standard deviations are similar, representativeness between the corpus as a whole and the measurable population can be claimed.

3.4.5 Criterion 3: Balance

A balanced corpus is achieved through intuition and accurate estimation (Clancy, 2012, p.86). One way to create a balanced corpus is to estimate the composition of the whole population of texts, then replicate the proportions of the categories of texts in the corpus. Alternative approaches include creating a corpus that consists

of sub-corpora containing equal number of texts or words. Oakey (2009) coined the term *isotextual approach* to describe collecting the same number of texts and *isolexical approach* to describe collecting the same number of words. If adopting an isotextual approach, the number of words will vary. One way to address any criticisms of this approach is to create a smaller balanced isolexical corpus and conduct a comparative statistical analysis to ascertain the degree of variation between the two approaches. Homogeneity is one aspect of balance that can be visualized using box and whisker plots to identify the sections of the corpus or population that may contain disproportionate numbers of outliers and unduly skew the results. The outliers should not, however, be duly excluded from the corpus, but the presence or absence of outliers should be noted when presenting results.

3.4.6 Criterion 4: Sampling frame

A sampling frame is the operationalized definition of the target population. Referring to sampling theory, Biber (1993) asserts that a clear definition of target population and method of sampling are key to achieving representativeness. There are no generally agreed criteria on sampling frames. Clear (1992) noted three difficulties when defining a sample:

1. the estimation of the population and therefore the calculation of an appropriate sample size,
2. the lack of a unit of a language to define the population (e.g. characters, letters, tokens, words, sentences, paragraphs, or texts), and
3. the uncertainty of accounting for all instances of language.

Once the appropriate measurement unit is selected, the ratio of the sample-population can be chosen. For very small populations, it may be possible to collect every text and so there the sample-population ratio is $1:1$. However, for the majority of corpora, this is an impossibility. In order to determine the size of a representative sample, it is necessary to know, or at least estimate, the size of the whole population of texts, and then calculate the minimum representative size of a sample (i.e. the corpus) that could be used to act as a proxy for the population. As noted in Subsection 3.4.4 on representativeness, population can be further operationalized as:

- the actual population, which may be difficult to quantify; and
- the measurable population, which can serve as a proxy for the actual population.

In social sciences random sampling tends to predominate, but Buchstaller and Khattab (2013, p.77) argue that in relatively homogenous populations (such as abstracts within a single discipline) a more tractable way is by using systematic sampling, such as the first ten texts of every hundred.

3.4.7 Section summary

Corpus selection involves a multitude of choices, each of which affects the corpus and, in turn, the results that can be achieved using the corpus. Regardless of the parameters and values chosen for different aspects, a key aspect is to provide sufficient detail about the corpus creation process so that third parties are able to make their own judgements about the quality of the corpus in terms of its suitability for purpose. For a specialist corpus of scientific research abstracts, the corpus size will be comparatively small. Since this project focuses on rhetorical moves, annotation will be necessary and so the size of the corpus will be limited by the time and financial cost of the annotation procedure. This is the focus of the following Section 3.5. The representativeness, balance and sampling frame of the corpus need to be selected, justified with reference to the literature and, where possible, statistically verified.

3.5 Corpus annotation

3.5.1 Overview

Subsection 3.5.2 begins by introducing manual and automatic annotation methods and describing two common types of items that are annotated: form and function. The level of description of this dissertation is pragmatic and so the main focus in the tagging of rhetorical moves. Section 3.5.3 discusses the concept and selection of ontological unit. Subsection 3.5.4 describes annotation procedures, such as schemas, protocols and training. A summary is given in Subsection 3.5.6

Section 3.5.5 discusses inter-annotator agreement measures in relation to this study. Section 3.5.6 provides a summary of the points related to corpus annotation that are most pertinent to this study.

3.5.2 Types of annotation

Raw corpora are not annotated and simply contain the texts that were collected, collated or crawled. Annotated corpora add another level of meaning in which the corpora are enriched with further linguistic information. To quote Zeldes (2018, p.6), "annotation is a consistent type of analysis, with its own guidelines for the assignment of values in individual cases".

It is common practice to keep the raw text and annotations separate and to use software to link the files together. Annotation of texts in a corpus are no longer stored within the same file as the original text. A text can be annotated multiple times. Each annotation forms one layer that can be visualized as an overlay that can be placed on top of the text so that the text is directly linked to its associated annotations. UAM CorpusTool (O'Donnell, 2008), for example, accesses the text files (.txt) and the annotation files (.idx) and links the data together through its interface. Annotation can be manual, automated or a combination thereof. Intra- and inter-reliability statistics

are usually calculated to quantify the level of agreement between or among coders to establish the degree reliability.

Manual annotation tends to be rather subjective and time-consuming (Dayrell et al., 2012). Human annotators can annotate texts and with reference to a predetermined protocol codified in a training manual and annotation guidelines. Hovy and Lavid (2010, p.19) states that the "stability of the annotation scheme" can be maintained by working closely with annotators and measuring inter-annotator reliability. Although the annotation is manual, human annotators still need to use some software to enable digital searching and marking-up of the annotations. There are a number of freely-available annotation tools, such as the UAM CorpusTool (O'Donnell, 2008).

3.5.3 Ontological unit

As described, rhetorical moves may, at times, be difficult to identify in context. This is because of the form-function dilemma. One of the primary decisions is the selection of the ontological unit to be annotated. In the case of rhetorical moves, there are two options: sentential or a string of words. Annotating at sentence level is far more straightforward since the boundary of each ontological unit is marked by an end stop. Shorter ontological units will result in lower degrees of inter-annotator reliability and greater annotation cost (in terms of either time or money). Longer ontological units will, however, loose granularity and may not provide a complete picture of the state of rhetorical moves, particularly embedded moves. Given that in the pilot study no embedding was discovered in the first five disciplines annotated (Blake, 2014), the selection of the sentence as the ontological unit appeared uncontroversial. The use of the sentence as the ontological unit is in line with Dayrell et al. (2012, p.1607) who noted "the vast majority of sentences from English abstracts reflect one single rhetorical move". Software, such as the UAM CorpusTool (O'Donnell, 2008) can automatically identify end stops and so can automatically demarcate ontological units. The ambiguity of full stops used in abbreviations, however, at times necessitates the annotator to refine the automatically identified ontological unit.

3.5.4 Annotation procedures

Leech (1993) proposed seven maxims for corpus annotation, which are reproduced below:

1. It should be possible to remove the annotation from an annotated corpus in order to revert to the raw corpus.
2. It should be possible to extract the annotations by themselves from the text.
3. The annotation scheme should be based on guidelines which are available to the end user.

4. It should be made clear how and by whom the annotation was carried out
5. The end user should be made aware that the corpus annotation is not infallible, but simply a potentially useful tool.
6. Annotation schemes should be based as far as possible on widely agreed and theory-neutral principles.
7. No annotation scheme has the *a priori* right to be considered as a standard.

Advances in corpus software since 1993 have addressed the first two maxims as by default the data (in read-only format) and the annotations are stored in different file formats. However, the remaining maxims are still applicable.

To quote Hovy and Lavid (2010, p.19):

> The more complex the phenomena being annotated, the more complex the theory generally is, and hence the more complex the instructions to the annotators.

Few studies have harnessed automated move annotation. There are two different approaches to automated annotation, namely a cluster approach and a linguistic feature approach. Different tools can focus on one of these approaches, or combine aspects of both. There are two notable automated annotation tools, namely: *Mover 1.0* (Anthony and Lashkia, 2003) and *MAZEA* (Dayrell et al., 2012).

One approach to automation utilizes a bag-of-words cluster approach in conjunction with n-grams and statistical method to automatically analyze the structure of a text (e.g. Pendar and Cotos, 2008). This approach underpinned the development of *Mover* (Anthony and Lashkia, 2003), a data-driven learning tool with the target audience being writers of technical research English whose first language is not English. Anthony and Lashkia (ibid., p.189) claim the "system performs consistently across different datasets with an average accuracy of 68%". This claim in not without contention, though. Some issues that they noted include insufficient training data for some steps, e.g. Step 1.1 and Step 2.1b. In their research study, sentences were classified as a particular move and a particular step within the move. There were a total of 3 moves and 12 steps. The most accurate classification was for Step 3.1b (announcing research) with an accurate ratio of 92%, but for Step 2.1b, the accuracy rate slumped to 17%. The training was conducted using the create-a-research-space framework (Swales, 1990) for introductions in computer science articles. However, whether the 68% accuracy can be achieved in other frameworks is yet to be determined. It should also be noted that *Mover* has not been updated since its release.

The other approach is linguistic, making use of various lexical, structural and syntactic features. Dayrell et al. (2012) created a natural language processing (NLP) pipeline called *MAZEA* that was designed to automate move detection in English research abstracts. A notable downside, however, for this software is the accuracy of

annotations of the training set and its consequential effect on the annotation of target sets.

Accuracy rates for automated annotation for both *Mover* and *MAZEA* did not rise above 80% and frequently resulted in significantly lower accuracy rates. If the rhetorical organization could be coded automatically with a high degree of accuracy, that would mean that a larger corpus can be analysed automatically. Should a reasonable degree of accuracy be achieved, semi-automatic coding could be adopted. Automated annotation appears to be no panacea for annotation projects. The selection of whether to use human annotators or automated annotation is contingent on the purpose of the research. Subjective judgements occur in both types of annotation with the subjective agent in automated annotation being the software developer while in manual annotation each annotator is a subjective agent. The moment the subjective decision is made, i.e. the subjective stage, also differs with the decisions being made prior to annotation in automated annotation while annotators make subjective decisions throughout coding. Automated annotation is highly replicable with perfect or near perfect scores since any subjective classification decisions were made during the development of the software and not on running the software. Manual annotation scores, however, vary according to annotators as annotators need to make classification decisions for each item.

One important factor in this project is the endogeneric nature of automated annotation. Texts are annotated based solely on the words occurring in those texts and the software does not draw upon any knowledge that is not encoded in the words in the text. Human annotators can and frequently do draw upon their world knowledge to use both endogeneric and exogeneric factors to make judgements. By drawing upon this knowledge, human annotators can make more accurate decisions about the coding of a particular move. However, when human annotators possess different levels of knowledge, codings may differ.

Form tagging, especially part-of-speech (POS) tagging, has a long history in corpus and computational linguistics. There are various part-of-speech (POS) taggers and tag sets. In general, it would seem prudent to select a tagger which can tag with the highest accuracy. Accuracy of POS-tagging is discussed in Subsection 3.5.5, and despite its relatively high accuracy of over 95%, the residual 5% error may cause false positive results.

In comparison with tagging form, tagging function is far more complex. Form, such as part of speech has a one-to-one or one-to-many relationship between word and part of speech and is inherent in the word. Function, however, is related to not only the choices of word or words, but the co-text, context and at times can only be deduced by using exogeneric knowledge of the content or drawing upon world knowledge. When annotating, one of the main decisions relates to the unit of meaning, i.e. the ontological unit.

Once the ontological unit is decided, labels need to be assigned to each rhetorical

move. The tagset can be predetermined, determined during annotation or a combination of both. Issues that arise when tags are amended or appended include the need to retag texts that had been annotated prior to the revision. One way to ameliorate the need for revision is through conducting a pilot project to identify the particular tags and tagset to be used for the main project.

Annotation guidelines are used to increase annotator agreement by providing clearly delineated categories, detailed protocols, numerous examples and discussion of boundary cases, creating a set of principles on which annotation decisions can be made. Even when only one annotator tags a corpus, there are problems of intra-annotator reliability. The codification of decisions and demarcation of boundaries of tags enables the annotator(s) to refer to a source rather than rely on their memories. This is particularly important in large-scale projects and those projects whose duration stretches across years.

The primary researcher is likely to be far more familiar with the intricacies of the texts to be annotated and the tagset to be used. A key issue is establishing an annotation procedure so that second annotators are able to make appropriate decisions based on the annotation protocol. This can be achieved through training, benchmarking and monitoring.

3.5.5 Inter-annotator agreement

Researchers cannot measure the correctness of annotations directly (Boleda and Evert, 2009), and so resort to reliability as a proxy variable. Reliability of annotations is evaluated through various on inter-annotator agreement (IAA) measures. Inter-annotator agreement or disagreement is a measure comparing the degree of agreement between two annotators or among more annotators (Artstein and Poesio, 2008). Perfect agreement results in a score of 100%, but more typically scores are markedly lower. According to Bayerl and Paul (2011), simple measures, such as observed or raw agreement measures are the most frequently used. These measures, however, are far from reliable. Yet, comparing the observed agreement alone does not take into account the expected percentage of agreement that would happen by chance. Therefore, to correct for this a statistical measure such as the kappa/alpha family (Artstein and Poesio, 2008), could be harnessed. Passonneau (2006, p.836) writes:

> "Measuring inter-annotator reliability involves more than a single number or single study. Di Eugenio and Glass (2004) argue that using multiple reliability metrics with different methods ... can be more revealing than a single metric. Passonneau *et al.* (2005) present a similar argument for the case of comparing different distance metrics."

Double annotation tends to be conducted for only part of a corpus. State-of-the-art annotation may be defined as annotation which is breaking new ground. Gold standard annotation is often mentioned. The gold standard of a corpus tends to be a

relatively small sub-corpus given the time and financial cost of human annotation (A. Roberts et al., 2007, p.626). Dayrell et al. (2012, p.1607) consider the gold standard for their automated annotation of abstracts as "the kappa between human annotators". which was calculated at sentence level, and based on human annotation of 72 abstracts (two groups of 38 and 34, a total of 5% of their corpus). Their Kappa statistics were 0.652 (N =306, k = 3, n = 20) and 0.535 (N = 148, k = 3, n = 18). These values show substantial agreement and moderate agreement, respectively.

The underlying, but flawed, assumptions are that lack of IAA rules out validity and high IAA implies validity.

Claims of annotation accuracy of over 95% are made for part-of-speech (POS) taggers (Gries and Berez, 2017). The Standford Tagger using maximum entropy cyclic dependency network (Gelbukh, 2011) assigns POS tags with an accuracy of 97.32%. This is close to 100% and so the 2.7% accuracy loss may be assumed to be insignificant. Yet, as Gelbukh (ibid., p.171) points out, this loss rapidly accumulates across a typical sentence. Assuming a sentence contains 21 words, and each word is tagged with an accuracy of 97.3%, the probability that the sentence is accurately tagged is slightly over 56%. Interestingly, Charniak (1997) points out that for a simplistic tagging method that assigns any word not in the training corpus as proper noun, an accuracy rate of 90% for POS tagging can be achieved.

Corpus annotation involves a trade-off between the "sophistication of categories and the practical attainability of a stable annotation" (Hovy and Lavid, 2010, p.19). Gelbukh (2011, p.171) also notes that this accuracy may drop further when the training and operational data differ in terms of "topic, epoch or writing style." Annotation should aim for both accuracy and efficiency. However, quoting Márquez and Giménez, "there is a trade-off between [these] two desirable properties. Greater accuracy invariably requires on processing more information, and so the key question is whether the accuracy is sufficient for the research" (ibid., p.171). In addition, Gelbukh (ibid., p.173) points out some other that factors affecting inter-annotator agreement for humans include task aptitude, degree of attention, level of guidance and degree of recall of guidance. Human annotators have the ability to draw upon exotextual factors and so an annotator with no specialist knowledge annotating computer science texts is likely to make different judgement calls than an annotator with specialist knowledge. Raghavan, Fosler-Lussier, and Lai (2012, p.1372) found that annotators with nursing background used "clinical judgment to infer certain annotations that were not directly observed in the [corpus] data when annotating texts for sets of words that indicate medical events. For instance, classifying something as an acute condition based on readings or values in the text."

The methodological choices aim to affect the judgments of the annotators in such a way that annotators make the same judgment call about which label to assign to a language item. Rissanen (1989, as cited in Archer, 2012, n.p.) points out the "mystery of vanishing reliability", i.e. the statistical unreliability of annotation that is too detailed. In short, the more tags, the less agreement. Although this may be

obvious with hindsight, researchers tend to develop tags that will inform the purpose of their research rather than manipulate a higher IAA by pragmatic choices of tags and tagsets. Such pragmatic choices might involve adopting catch-all tags rather than finely-nuanced tags. Some methodological choices that enhance IAA include effective training, use of annotation guidelines, size of tagset and clarity of tag demarcation.

3.5.6 Section summary

Corpus annotation is a complex task that requires thorough preparation. Key concepts include the choice of the ontological unit and the desired level of granularity, which affect the tagset to be used. To ensure annotation accuracy, an annotation guide should be created to enable human annotators to benchmark their choices against a guiding set of principles. Where double annotators are used, IAA scores can be calculated to evaluate the degree of discrepancy between judgement calls. When IAA scores are harmonious, automated annotation is more likely to be possible, and when IAA scores are lower, it is likely that there are confounding factors at play that affect human annotators differently. There are many possible causes of low IAA scores, including differences in the annotator's knowledge base and their ability to recall the annotation guidelines.

For a corpus of scientific research abstracts, there does not appear to be a viable way of automatic annotation, and so human annotators need to be used to annotate the rhetorical moves. When the judgement call is based on the annotation guidelines, a top-down annotation approach is being used, but when annotation guidelines are altered based on the corpus, a bottom-up approach is being used. The most likely approach is one that starts top-down and is refined in a cyclical manner during the early stages of the annotation process.

3.6 Chapter Summary

The review of the literature on corpus linguistics raises some cogent arguments for the adoption of a corpus linguistic approach to this study. In short, corpus linguistics is a methodology that enables empirical quantitative research to be conducted on texts using descriptive and/or analytical statistics.

To systematically investigate the language used in research abstracts, it is not sufficient to rely on intuition and recollection, and so a corpus becomes essential. In short, empirical evidence for a statistical study of language can be obtained through descriptive analysis of a suitable corpus. In this study, a specialist corpus containing research abstracts from the target publications provides better quality data than a more general larger corpus.

Adopting a corpus linguistic methodology allows texts to be read in non-linear manner; thus, enabling data to be extracted that traditional linear reading may not have revealed. The annotations add value to the text by enabling researchers to

extract particular categories (such as moves and sub-moves) and identify both when and where they are used in relation to other categories. In addition, the annotations enable researchers to pinpoint, compare and contrast language used to realize those specific moves and sub-moves within and among scientific disciplines.

Rhetorical moves in scientific research abstracts can, therefore, be investigated in a systematic, replicable manner. This can provide a firm foundation to investigate the lexical realization within the rhetorical moves.

At the outset of this study, to the best of my knowledge, there was no freely-available balanced representative corpus of research abstracts that covered a wide range of scientific disciplines with a focus on the hard sciences, and so a tailor-made corpus was created. To ensure that a corpus is representative, balanced and suitable for the purpose, the sampling frame and sampling unit need to be established. Saturation and representativeness ought to be measured and evaluated.

Chapter 4

Methodology

> *Measurement has too often been the leitmotif of many investigations rather than the experimental examination of hypotheses. Mounds of data are collected, which are statistically decorous and methodologically unimpeachable, but conclusions are often trivial and rarely useful in decision making. This results from an overly rigorous control of an insignificant variable and a widespread deficiency in the framing of pertinent questions. Investigators seem to have settled for what is measurable instead of measuring what they would really like to know.*
> - Edmund D. Pellegrino, 1964, Medical doctor and academic

4.1 Chapter preview

This chapter describes the research method selected to collect, process and analyze the data needed to answer the research questions arising from the literature review (2). The previous chapter on corpus linguistics (3) provides the foundation on which this research method is based. This chapter is divided into four sections, namely: a broad overview of the three-phase research process followed by sections on the corpus, annotation and analysis phases.

Section 4.2 begins by providing an overview of the research process used and the sequencing of the three-phase method. It then provides a non-technical explanation of the research method adopted to address each research question. The section concludes by detailing a stance on research and on the importance of data quality.

Section 4.3 describes the corpus phase in depth. Descriptions of the corpus selection, corpus specification, corpus collection and cleaning stages within this phase are provided. The standard operating procedures (SOPs) created for each stage are described.

Section 4.4 details the steps taken in the annotation phase. An overview of the pilot studies and trials which were carried out prior to the actual annotation is given. Explanations of how these pilots and trials informed the annotation phase are given in the relevant subsections. The annotation phase includes descriptions of the annotation protocol codified in the detailed annotation guide specifically created for this project. The recruitment, training and benchmarking procedure for annotators is

explained, and measures to evaluate inter-annotator reliability and decisions on their interpretation are provided.

Section 4.5 describes the analysis used on the annotated corpus to answer the research questions. The standard operating procedures to extract and analyze the required data points and datasets for each sub research question are detailed to enable the replication of this study.

A chapter summary is provided in the final section (Section 4.6).

4.2 Research process

> *Research is formalized curiosity. It is poking and prying with a purpose*
> - Zora Neale Hurston, American anthropologist

This section starts by introducing the three-phase process used in this study (Subsection 4.2.1). This is followed by an easy-to-follow quick guide to the specific steps involved in answering each main research question in Subsection 4.2.2. A declaration of assumptions is provided in Subsection 4.2.3. The section ends with an explanation of how data quality was managed in Subsection 4.2.4.

4.2.1 Three phases in the research process

Figure 4.1 shows the three phases of this research process. The corpus collection phase involves the selection, specification, collection and cleaning of a corpus. The annotation phase can be divided into five stages: preparatory trials, the development of the annotation protocol, annotation of the whole corpus, annotator training, double annotation and determination of the inter-annotator reliability. The final phase of analysis has three components. The first is the pre-processing of the datasets while each of the remaining two stages addresses one of the research questions. The analysis of rhetorical structure necessarily precedes analysis of lexical realization within the rhetorical moves.

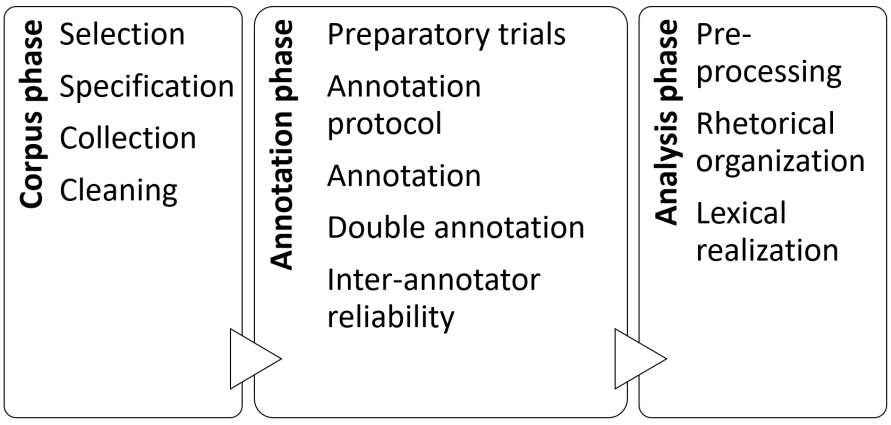


FIGURE 4.1: Three phases

4.2.2 Research method by Research Question

This subsection provides a relatively non-technical overview of the steps involved in answering the research questions (2.8).

Research method for Research Question 1

To understand the rhetorical organization of abstracts of research articles published in a broad range of top-tier scientific journals, the following actions were undertaken.

1. Created a corpus of scientific research abstracts.
2. Labelled the moves occurring in each research abstract using human annotators.
3. Verified annotation accuracy.
4. Extracted the sequence of move labels in each abstract using a tailor-made script.
5. Counted the permutations of move sequences in each discipline.
6. Plotted, compared and contrasted the rhetorical organization by discipline.
7. Evaluated values for linguistic dimensions.
8. Applied multidimensional scaling and cluster analysis to the linguistic dimensions to identify the similarities and differences in rhetorical organization between the disciplines.

Research method for Research Question 2

This research question focuses on lexical features. The selected lexical features are keyness (e.g. key word lists) and grammatical tense (e.g. verb forms that carry tense and may show aspect). The method to discover the lexical features of prototypical moves in abstracts of research articles in the selected scientific disciplines is stated below.

1. The key words for the sub-corpora of individual moves in each discipline are determined.
2. The grammatical tense of finite verbs within each move and each discipline is identified and counted using a tailor-made script.
3. The frequency of key words are compared and contrasted by move and discipline.
4. The frequency of grammatical tenses are compared and contrasted by move and discipline.

4.2.3 Declaration of assumptions

Researcher bias

Assumptions are omnipresent in linguistic analyses. This declaration aims at "taming' one's subjectivity" (Peshkin, 1988, p.20) by stating from the outset the initial (known) assumptions and acknowledging the likelihood of unknown assumptions. Baker (2006b, p.92) astutely points out that researcher bias can account for both the discoveries made and the discoveries missed. Researcher bias cannot be avoided, but the usage of corpora and quantitative analysis helps ameliorate this bias by providing quantifiable values on which research decisions can be based.

A priori expectations

Having conducted research in linguistics, education and management, I was familiar with research abstracts in three research disciplines. My expectation was that research abstracts provided key details of the research and served as a taster to help readers assess whether the full research article is worth reading. This "worth" may relate to the relevance of the research area and/or the results of the research. I expected research abstracts to be relatively short self-contained stretches of discourse that present the research in a linear manner.

At the start of this research, my knowledge of the technical terminology used in disciplines such as information theory and image processing was very limited. I relied on deducing meanings from the lay meaning of words and using contextual clues.

Subsequent stance

Through conversations with subject specialists I became aware, however, of the gulf between my understanding of an abstract and that of the subject specialist. To narrow the gap I read the core university texts for hard science disciplines, particularly focusing on the applied sciences as those abstracts were the most difficult to extract meaning from. The five disciplines that were more difficult to understand were information theory (IT), wireless computing (WC), image processing (IP), evolutionary computation (EC), and knowledge and data engineering (KDE). This background reading and knowledge provide me a firmer grounding in the concepts and terminology. Although far from a subject specialist, greater familiarity with technical terms, schemes and concepts makes it much easier to understand the intended meaning of the writers of scientific abstracts in these disciplines.

4.2.4 Data quality management

A fundamental concept in computer science is "garbage in, garbage out". This adage also applies to research using corpus linguistics (Brezina, 2018, p.262; Brysbaert, Mandera, and Keuleers, 2017; Geiger et al., 2020; Hardie, 2007, pp.23–24; McEnery

and Hardie, 2012, chapter 2; Mindt, 2002, p.198). Should the results obtained be based on flawed data, the results would be invalid. Data processing can be viewed as a five-step process (De Jonge and Loo, 2013) as shown in figure 4.2.

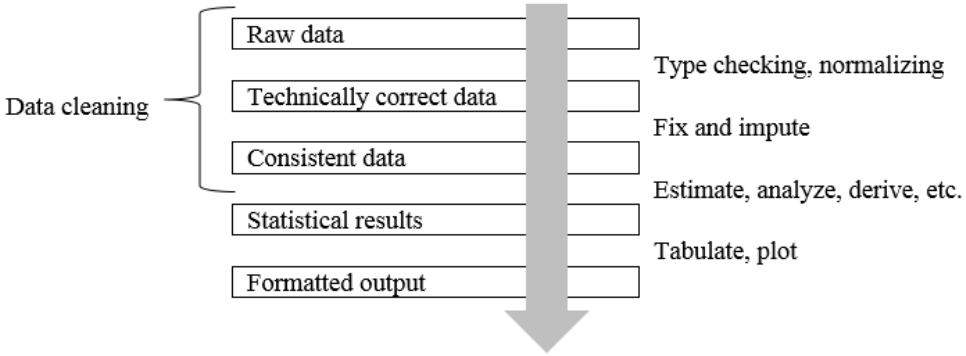


FIGURE 4.2: Data value chain
Source: De Jonge and Loo (2013)

De Jonge and Loo (ibid.) classify data into three levels: raw data, technically correct data and consistent data. Raw data is the data that is collected or received. Files may contain wrong data types, wrong labels, character encoding problems, etc. This data needs preprocessing to convert into a state that can be input into statistical software, which in the case of R is into an R data.frame. Technically correct data is that which meets the minimum criteria for input into an R data.frame, and so contains the correct names, types and labels. However, despite its technical accuracy, there are still likely to be a number of errors that will affect the quality of the results, such as missing or corrupted data that need to be addressed. Kimball and Caserta (2004, p.15) suggest replacing missing data with a "question mark or supplying least biased estimators of numeric values". Data quality processing involves multiple steps, such as checking the consistency of values, removing duplicated values and, at times, judgement calls on when human intervention is necessary to ensure that the data conforms to requirements (ibid., p.18). The third level of data, namely consistent data, is data from which valid statistical inferences can be made. The statistical results from this consistent data can then be tabulated or plotted.

4.3 Corpus phase

4.3.1 Overview

The corpus phase comprises four stages, namely corpus selection, corpus specification, corpus collection and corpus cleaning. Subsection 4.3.2 describes the choices made at the level of discipline, publication and text type. The corpus specification including the use of saturation to estimate the corpus size is given in Subsection 4.3.3. The corpus collection procedure is described in 4.3.4. The corpus cleaning procedure is the focus of 4.3.5.

4.3.2 Corpus selection

An analytic hierarchy decision-making process (Mu and Pereyra-Rojas, 2017) was used to select the specific disciplines, publications and text types within those publications. The decisions with their rationales are detailed in this subsection.

Discipline selection

To enable the widest possible generalization within the exigencies of time and financial resources, a taxonomy of scientific disciplines was mapped, and specific disciplines selected from each branch of the taxonomy with a particular focus on those disciplines that are under-represented in the research on scientific abstracts.

Demarcation of science and non-science disciplines is problematic with philosophers having attempted to do so for millennia (Resnik, 2000). The demarcation of disciplines within science is no less problematic. Both classification dilemmas are forms of the lack-of-boundaries version of Sorites paradox (Oms and Zardini, 2019).

Becher and Trowler (2001) divide disciplines into a two-by-two matrix as shown in Table 4.1. This matrix classifies disciplines into one of four quadrants: hard pure, soft pure, hard applied or soft applied sciences. This, as Becher and Trowler (ibid.) note, is an over-simplification, but it provides a roughly-grained classification onto which disciplines can be mapped although given the fuzzy borders of some disciplines, they may not fit perfectly within one quadrant. As noted by Becher and Trowler (ibid.) and Harwood (2005), this taxonomy does not account for the numerous disciplinary differences. Put simply, hard sciences could be described as those that are governed by laws while soft sciences are those that are commonly referred to as social sciences. Pure sciences are more theoretical and focus on describing the world. Applied sciences are more practical, and focus on real-world applications that change the world. The distinction between pure and applied science is also argued to relate to the "commercialization of scientific knowledge" (Lucier, 2012, p.249).

TABLE 4.1: Becher taxonomy

	Hard sciences	**Soft sciences**
Pure sciences	e.g. Physics	e.g. Economics
Applied sciences	e.g. Computer science	e.g. Business & Management

Source: Becher and Trowler (2001)

Although this classification uses the term "soft sciences", scientists who believe that only disciplines with laws (as opposed to rules) are science, may see "social science" or "soft science"as euphemisms for humanities. Both social science and humanities are used to describe research on society and human relationship as a whole. The pure sciences include natural, social, and formal science. Both natural and social sciences are empirical sciences (Venable, 2006, p.2). This gives rise to the taxonomy as shown in figure 4.3.

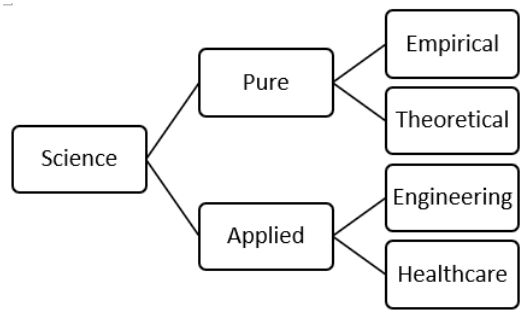


FIGURE 4.3: Initial discipline taxonomy

Natural science can be subdivided into physical science and life science. Formal science includes subjects such as logic, mathematics and theoretical computer science, all of which are theoretical, focussing on laws and rules. The pure sciences provide the foundation for applied sciences, which can be broadly divided into two categories: engineering and healthcare. Interdisciplinary sciences, as the name suggests, combine knowledge from two or more disciplines. Starting from Becher's taxonomy (Becher and Trowler, 2001) and incorporating some of the finer divisions described above, a six-branch discipline taxomony was created. This taxonomy provides a more finely-grained taxonomy than Becher and Trowler (ibid.) with less emphasis on social sciences. Figure 4.4 shows a taxonomy that moves away from the soft/hard distinction. This taxonomy divides science into six branches. The social sciences and healthcare sciences are well represented in the research literature while there is very little research on engineering.

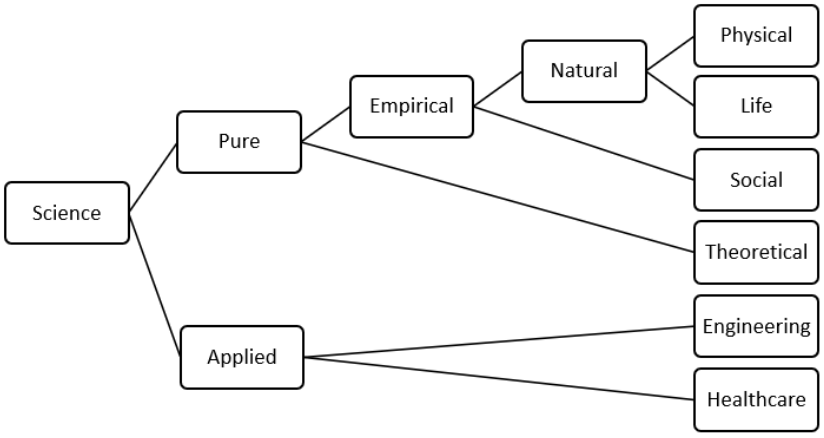


FIGURE 4.4: Extended discipline taxonomy

The disciplines were selected using two criteria. The first criterion was to choose disciplines for which access to specialist informants was available. The importance of access to subject specialists has been documented by numerous authors (see for example, Bhatia, 1993; Huckin and Olsen, 1984; Selinker, 1979; Swales, 1990; Hyland, 2005a). Hyland (2005a) notes that specialist informants can help confirm and validate findings. The second criterion was to ensure a range of scientific disciplines were

represented with a particular focus on under-researched disciplines. As there is a paucity of corpus research on engineering, particularly those within the multidisciplinary field of computer and information science, the number of disciplines selected for engineering was increased. The remaining choices of discipline were made to ensure that each branch in Figure 4.4 was represented by at least one discipline. This results in a corpus divided into ten disciplines as shown in Figure 4.5.

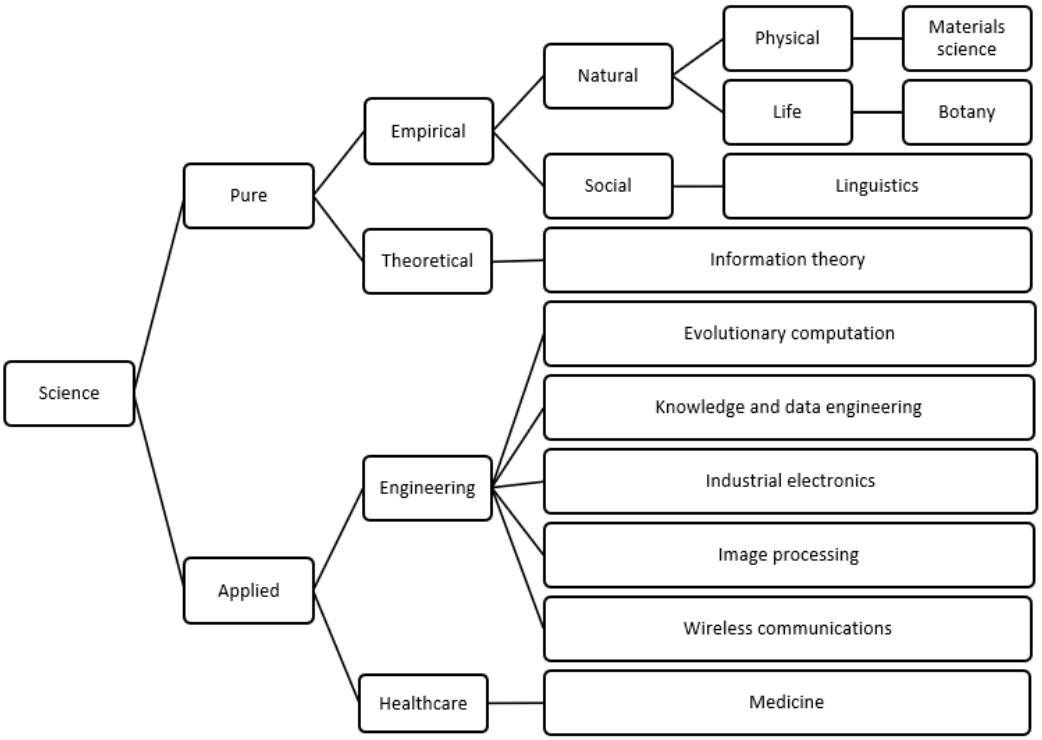


FIGURE 4.5: Final discipline taxonomy

In summary, ten different academic disciplines were chosen. One discipline each was chosen for the sub-branches of physical science, life science, social science formal science and healthcare. Five disciplines were chosen for engineering to address the gap in the research literature. It should be noted that each of these disciplines continues to furcate, for example within linguistics, there is theoretical, forensic and applied linguistics, etc.

Publication selection

The two most common types of research publications are conference proceedings and journals. Both forms of publication vary in quality from top-tier to bottom-shelf (for a discussion on criteria on which publications can be measured see Gu and Blackmore, 2017). In most disciplines journals are more prestigious due to their more stringent peer review process. Computer science is a notable exception (Kim, 2019). As one of the primary aims of this research is to help novice writers understand the genre of research abstracts, top-tier journals were selected. There are two reasons for this choice. First, any findings based on top-tier journal articles may be generalized to

lower-tier journals. Authors usually aim high, but tend to resubmit papers that are rejected from higher-tier journals to lower tiers (McDonald, Cloft, and Kallmes, 2007; Ray, Berkwits, and Davidoff, 2000). Second, most researchers aspire to publish in these types of journals as the benefit of publishing one paper in a top-tier journal outweighs numerous publications in lower-tier journals (Reich, 2013).

The selection of specific journals was based on three criteria: accessibility, ranking and expert recommendation. Access to the full text was deemed necessary to be able to clarify the meaning of particular sentences when the lack of contextual information combined with the lack of disciplinary knowledge made classification difficult. Second, journals that were highly ranked as as judged by h-index factors (Hirsch, 2010) were chosen. It should be noted that there are criticisms on the veracity and usefulness of impact factors (Chapman et al., 2019; Tort, Targino, and Amaral, 2012). Bollen, Rodriquez, and Van de Sompel (2006) note that like people, some journals are popular but not prestigious and vice versa. Citation indices reflect the number of citations and, as such, can be used as a measure of popularity. Bollen, Rodriquez, and Van de Sompel (ibid., p.670) also point out that review articles may contain no cutting-edge research, be highly cited by graduate students but ignored by experts. Third, journals that specialist informants recommended as appropriate publication venues for themselves or their doctoral students were given priority.

Text type selection

Scholarly journals publish not only research articles but a number of other types of articles, such as book and conference reviews. This study focuses on abstracts of research articles and so abstracts of other text types were not collected. Different journals, however, have different naming conventions for research articles and different expectations regarding the length of articles and their associated abstracts. What may be classed as a short research article in one domain could be considered as a full research article in another domain. Given this, abstracts from all research articles regardless of length of the accompanying research article were selected. Specific details of the text types are detailed below:

IEEE journals Engineering research articles in the *Institute of Electrical and Electronics Engineers (IEEE) Xplore* collection of journals are published under various names. Full-length articles tend to be named articles, while short papers may be classed as research communications or letters. Abstracts from articles, research communications and letters were selected for inclusion.

Linguistics journals In the *Journal of Communication* (Volume 62) abstracts of articles classed as original articles were included while book reviews were excluded. In the journal *Applied Linguistics* abstracts of articles classed as either articles or forums were included while reviews were excluded. In the *Journal of Cognitive Neuroscience (Volume*

24) there was only one class of article, so abstracts of each article were selected for inclusion.

Medical journal Abstracts of articles in the *British Medical Journal (International edition)* classed as research were selected for inclusion.

Botany journal Abstracts of articles classed as research articles and large-scale biology articles in *The Plant Cell* were included. Given that the primary practical purpose of this research is to help novice writers and the fact that review articles are usually written by experienced authors, article classed as reviews were excluded.

Materials science journal Abstracts of full research articles and those published in the communications section of *Advanced Materials* were included.

4.3.3 Corpus specification

The texts to be collected have been identified, but the number of texts necessary to answer the research questions and provide a representative sample needs to be determined. As there is a substantial time cost in annotation, a corpus that is sufficiently representative is optimal. Annotation of dense, terminology-laden texts on obscure areas of technical research substantially increases the difficulty of coding texts. With this in mind, the ideal scenario is to identify the minimum number of texts that would be needed to achieve the aims of this research with the degree of accuracy required. Given that the importance of moves, moves were used to assess the required corpus size.

Following the advice in McEnery, Xiao, and Tono (2006, p.16) closure or saturation was determined in an early pilot study by comparing the yield of new items in subsequent corpus segments with previous corpus segments. A script in C was created to extract and count the combinations of rhetorical moves (see Appendix A.3 for the C script). For hypothesis-testing research, one way of calculating the optimal size of a corpus may be determined by plotting the graph of the accumulated variation in the particular features of interest against the number of texts or ontological units in the preliminary phase of corpus creation. The law of diminishing returns can be harnessed to ascertain the point of saturation (Belica, 1996) or closure. Long-tail distributions at word level complicate pinpointing saturation, but at the level of rhetorical move, this is a moot point. A cost-benefit optimization can be used to determine the cut-off point for the optimal size of the corpus. A graph of cumulative and projected rhetorical move combinations against the number of abstracts annotated was plotted. Figure 4.6 shows the graph of diminishing returns created during a trial study (Blake, 2015b). It can be clearly seen that in the first twenty-abstract increment there is a high number of new permutations identified, yet in subsequent increments of twenty abstracts, the number of new combinations identified decreases. Assuming that this trend continues, the graph will level off at some point.

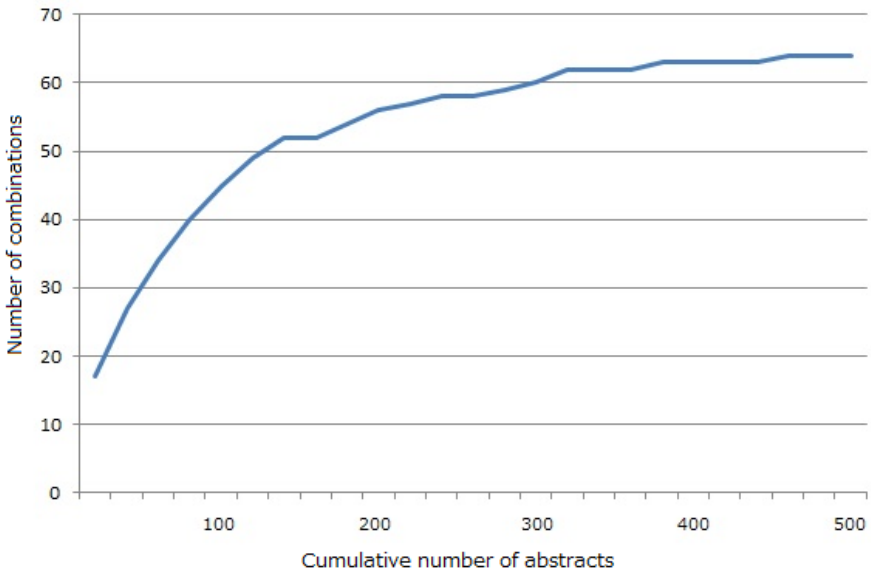


FIGURE 4.6: Graph of diminishing returns

Based on the pilot study, it was determined that 100 abstracts per discipline would suffice. One hundred abstracts per discipline was also the size chosen by Hyland (2004).

An isotextual approach (Oakey, 2009) was adopted, in which the same number of abstracts were collected for each discipline regardless of the length of the texts or the number of moves contained within the texts. In a pilot trial of 100 abstracts (20 abstracts per discipline), little difference was found in the number of words in each domain (Blake, 2014). Thus, given the similarity in both number and length of texts, there was no evidence to suggest that the adoption of an isotextual approach was contentious. When the corpus was extended to include abstracts from materials science and medicine, the variation in number of words became notable.

In line with the advice disseminated by Bender and Friedman (2018), data statements were created for this corpus under the headings of curation rationale, language variety, annotator demographics and text characteristics.

4.3.4 Corpus collection procedure

The contents of the corpus were chosen regardless of the particular linguistic features contained within the texts (Sinclair, 2004a). The corpus collection stage did not involve any linguistic analysis and simply involved the copying of electronic documents from online journals and conference proceedings and pasting the wording in individual text files which were named, saved and logged. Each abstract was stored as a separate text file using UTF-8 variable width character encoding with backups stored online, offline and offsite.

Given that the terms of use for institutional subscribers of IEEE Xplore specifically stated "Institutional subscribers are NOT permitted to . . . use robots or intelligent

agents to access, search and/or systematically download any portion of IEEE Xplore", each abstract was collected manually rather than creating a script to scrape the content. A standard operating procedure (SOP) was created to error-proof the process. The number of construction criteria were kept to a minimum in line with the basic principles advocated in Sinclair (2004a). The criteria for the collection of the texts was to collect abstracts of the pre-determined text type for each selected journal, starting from the first issue published in 2012 and continuing until the required number of texts were collected.

To minimize the possibility of human error and enable replication an SOP was created for the collection procedure. Ideally, this could have been automated with a script, but due to copyright restrictions, systematic automatic downloading was not possible. The complete SOP for corpus collection is given Appendix A.19.

Abstracts were collected from one or more top-tier journals for each discipline. Academic journals are issued at different frequencies with the most common being weekly, monthly, bimonthly or quarterly. Disciplines with larger numbers of researchers unsurprising have larger volumes of publications. A case in point is *Applied Linguistics*, a top-tier linguistic journal that publishes around 20 articles per year, while its equivalent journal in information theory, *IEEE Transactions on Information Theory*, publishes around 500 articles a year. As *Applied Linguistics* only published around 20 research articles in the year of collection, supplementary journals were added to make up the shortfall.

4.3.5 Corpus cleaning

> *If you think your data is clean, you haven't looked at it hard enough.*
> - Eben Hewitt, Author of Technology Strategy Patterns

The purpose of this research is to focus on the rhetorical organization and lexical realization of moves in scientific research articles. To be able to achieve this with the highest degree of accuracy, it is necessary to take steps to ensure that the corpus of texts files is suitable (Gries and Newman, 2013, p.264). Decisions on how to address factors that may negatively affect the results were taken for various elements, including typographic errors, random characters that appeared as a result of encoding errors, or collection errors, such as duplicate or incomplete abstracts.

One duplicate abstract was discovered. This was not a collection error. One journal had published the same article twice – once as the final article in an issue and then as the first article in the subsequent issue. Both articles were kept in the corpus.

Two key decisions were related to typographical errors and nonsensical characters. Typographical errors were not corrected. Typographical errors do not affect rhetorical organization at all and so have no impact on the main thrust of this research. Although typographical errors may affect the exact frequency counts for tokens in calculating key words when addressing lexical realization, the effect was felt to be negligible and so no alterations were made. This also avoided the necessity for demarcation

between errors that are typographical and those that are not. For example, we can assume that *hte* is *the*, since the word *hte* does not exist, and so that error is *typographic*. However, some spelling and usage errors are caused by ignorance, or the use of non-Anglophonic norms within scientific writing.

Spurious nonsensical characters that were clearly not intended to be part of abstracts and odd non-English characters within English abstracts were deleted. These appeared to be either remnants of LaTeX code or non UTF-8 characters. One annotator-introduced error was also corrected in which a sentence was coded twice: once including the end stop and once without the end stop.

4.4 Annotation phase

4.4.1 Overview

The annotation phase is concerned with the manual identification of rhetorical moves. The annotation approach could be called functional-semantic rather than linguistic as cognitive judgments were necessary to evaluate the intended meaning of each sentence given the cotext and content. This phase can be divided into three stages. In the preparatory stage (4.4.2) a series of pilot and trial studies were conducted. The results of these studies informed the choices of annotation tool, tagset labels, annotation schema, annotation protocol and content of training course. In the annotation stage (4.4.3) the whole corpus was annotated by the author. The accuracy and completeness of the annotations were verified by the author, and their veracity was confirmed by specialist informants. The author check was conducted using a specially-created online visualization tool, the *Move Highlighter*. The specialist informants accessed the annotated sub-corpora using a different specially-created tool, the *Move Visualizer*. In the final double annotation stage (4.4.4) double annotators were recruited, trained and benchmarked. After the double annotation was completed, inter-annotation agreement measures were calculated and interpreted.

4.4.2 Preparatory stage

Pilots and trials

There were numerous iterations of exploration and experimentation cycles (Wallis and G. Nelson, 2001). Extensive pilot testing and trials were completed. Based on the testing and trials, the most suitable ontological unit, tagset, schemata and annotation protocol were developed. Table 4.2 lists the pilot annotations and their respective effects on this study.

Details of pilot studies that were particularly notable are given below. Six of the pilot studies were delivered as posters or papers at international conferences to gain advice on how to improve the method.

TABLE 4.2: Pilot studies and trials informing this study

No.	Year	Corpus size[a]	Details	Effect on study
1	2012	40	Annotation using IMRD	Discrepency between views of specialist informants and linguists
2	2012[b]	134	Annotation using SFL (process, participant and circumstance)	Subjectivity of human judgment for borderline cases
3	2013[c]	5	Multimethod analysis	Awareness of cyclicity dimension
4	2013[d]	500	Analysis of research article titles	Awareness of disciplinary variation
5	2013	50	Ontological unit selection	Selection of sentence as unit
6	2014[e]	100	Tagset selection (CARS vs IMRD)	Selection of IPMRD tagset
7	2014[f]	500	Grammatical features	Paucity of discourse markers
8	2015	500	Determination of closure	Selection of 100 abstracts per subcorpus
9	2015[g]	500	Prescriptive-description disjuncture	Identification of linearity and variation dimensions
10	2016	2000	Automated annotation	Exogeneric knowledge essential for annotation accuracy. Selection of human annotation.

[a] Corpus size is determined by the number of abstracts

[b] Blake, J. (2012, December 7-9). Research abstracts: A diachronic systemic functional analysis. Paper presented at the 2nd GDUFS Forum on Applied Linguistics, Guang Dong University of Foreign Studies, Guangzhou, China.

[c] Blake, J. (2013, March 15-16). A multimethod analysis of scientific research abstracts. Paper presented at 3rd International Conference on Foreign Language Learning and Teaching, 2013. Bangkok, Thailand.

[d] Blake, J. (2013, November 30-December 1). A corpus-based study of scientific research article titles. Paper presented at Oita Text Forum Workshop 5, Oita University, Japan.

[e] Blake, J. (2014, March 7-9). Move structure of scientific research abstracts: CARS vs. IMRAD. Paper presented at the Second Asia Pacific Corpus Linguistics Conference, Hong Kong Polytechnic University, Hong Kong, China.

[f] Blake, J. (2014, June 20-21). Contrastive discourse markers in scientific research abstracts. Paper presented at 1st TRI-ELE International Conference on English Language Education, Bangkok, Thailand.

[g] Blake, J. (2015, July 20-24). Prescriptive-descriptive disjuncture: Rhetorical organisation of research abstracts in information science. Poster presented at the 8th International Corpus Linguistics Conference. Lancaster University, England.

Pilot study with specialist informants In this pilot study, five specialist informants annotated print-outs of abstracts using highlighter pens. Each abstract was printed out in large font on a single sheet of paper. The informants were provided with the tag set (INTRODUCTION, PURPOSE, METHOD, RESULT and DISCUSSION MOVE) and their definitions. They were asked to highlight when moves started and finished.

Two specialist informants who were full professors ignored the guidelines. The first professor created a new tagset of three labels: *novelty, substance* and *importance*. The other professor labelled each move as METHOD MOVE. His justification was that the background information was about a method. The purpose was to develop a method. The method was developed and then a method was used to evaluate the method. The three other specialist informants annotated in line with expectations and provided useful explanations of their decision process.

A key learning point was that without the background knowledge of the research domain, it would be difficult to differentiate between a results move and background information. Another problem was number of unknown technical terms, obfuscating the meaning. To ameliorate this terminology gap, I read the core set introductory textbooks for undergraduates for each of the five technical disciplines.

Pilot study with double annotators The training aid for this pilot study was a coding booklet which was in the main very similar to the final version of the booklet. The annotators used the UAM CorpusTool. Five volunteer annotators holding at least a master's degree in linguistics or a closely-related field were asked to annotate ten abstracts from each discipline. However, this proved to be time-consuming. It was necessary to show annotators how to use the tool as official documentation was difficult to follow. The lexical density and technical terminology, however, presented an insurmountable barrier to many of the annotators. Three annotators completed a small subset and gave up because they felt unable to complete the task. Two of the five annotators annotated the full set of 100 abstracts. However, the annotators held little confidence in many of their choices and frequently used the temporary label of "unknown" as the final classification. One annotator commented: "I read the [information theory] abstract multiple times, used Google, but in the end all I could work out was the part of speech of many words". Another annotator who held degrees in science, education and linguistics annotated abstracts found IT abstracts more challenging than envisaged as shown by his follow-up email:

> Five abstracts into the set, I have not yet found one that I can confidently complete. In four out of five, the entire discourse reads like background, or mere description of what is addressed. Any reference to "results" promotes the method rather than summarising findings. I can't discussion [sic] between specific "results as method" and broader "discussion". (Personal communication)

It was clear that without sufficient background knowledge, it would not be possible to annotate abstracts in the more technical disciplines of Evolutionary Computation (EC), Knowledge and Data Engineering (KDE), Image processing (IP), Information theory (IT) and Wireless communications (WC). Information theory proved to be the most challenging. The study showed the necessity for disciplinary expertise to break through the terminology barrier, and the necessity to provide more guidance on using the UAM CorpusTool.

Preparatory steps

These pilot studies and trial studies informed the development and refinement of the annotation process. The specific development steps are described in turn.

Step 1: Tool selection

The UAM CorpusTool (O'Donnell, 2008) was selected as the annotation tool because of its ability to create tailor-made annotation layers. The other viable options were GATE developer (Cunningham, Maynard, and Bontcheva, 2011) and Webanno (Yimam et al., 2013) both of which required setting up a server. The UAM CorpusTool could be downloaded and installed easily. One drawback was the limited how-to documentation for the tool.

Step 2: Ontological unit selection

Following the argument summarized in the previous chapter (3.5.3) and following numerous other researchers (dos Santos, 1996; Holmes, 1997) the ontological unit selected is the sentence. This decision was tested during a pilot study in which moves where annotated using two systems, namely IMRD and CARS (Blake, 2014). Few instances of embedded moves were discovered and so the decision to annotate at sentence level was confirmed.

Following Ren and Y. Li (2011) the length (measured in number of words) is concomitant with the importance that the author assigns to a move. Thus, when two moves are present in one ontological unit, the unit is coded according to the move with more words. In the event that the number of words is identical, based on the principle of information focus (Blake, 2015a), the first move is selected.

Step 3: Annotation labels (tagset) development

The tagset was informed by the research literature and developed through pilot studies (Blake, 2014; Blake, 2015b). In the finalized tagset, there are five main categories and one temporary category "unknown". The five main categories are: INTRODUCTION MOVE, PURPOSE MOVE, METHOD MOVE, RESULT (result or product) and DISCUSSION MOVE (discussion or conclusion). The temporary category is a way to mark up sentences that are not easily assigned to one or more of the five categories.

Sentences classed in "unknown" category are re-considered until a more suitable category is determined. There are five minor categories, namely BACKGROUND, PROBLEM, GAP, OVERVIEW and METHOD AS PRODUCT. The major categories are classed as moves, while the minor categories are sub-moves.

Step 4: Schema development

The annotation schema 4.7 is a tailor-made tagset that specifies the hierarchy of tags to markup text in an efficient manner. When this schema is specified in the UAM CorpusTool, the annotator is shown the choice of annotation labels according to the schema. Initial annotation choices are made from *MOVE TYPE*. If the labels <*introduction*> or <*result*> are selected, a second level of hierarchy then appears (*INTRODUCTION TYPE* or *RESULT TYPE*). The annotator then selects from the choice given. Once a sentence has been fully tagged, the next ontological unit is automatically selected. The schema was developed and refined until an optimal schema was finalized.

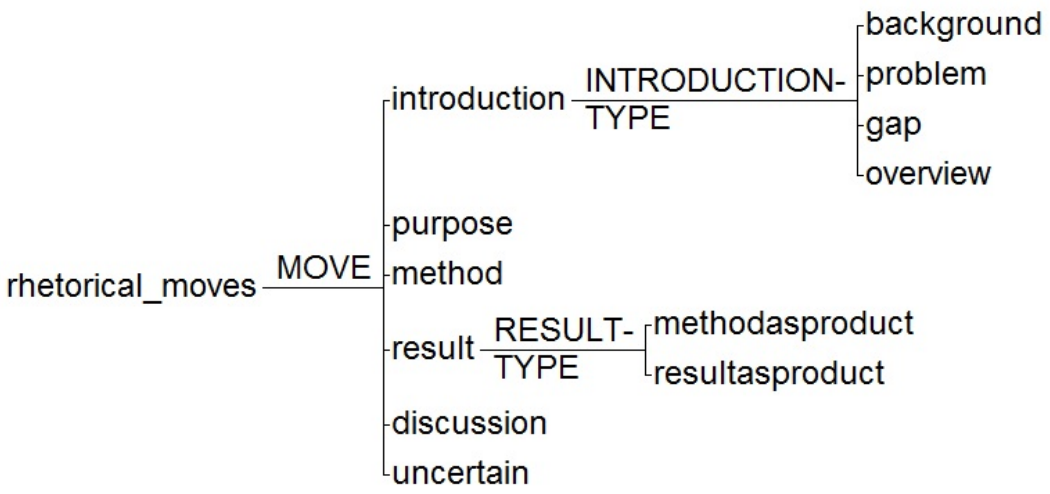


FIGURE 4.7: Annotation schema for move layer in UAM CorpusTool

Step 5: Protocol development

The creation of the annotation protocol is a non-trivial task (Hovy and Lavid, 2010, p.19). Wallis (2003, p.63) points out two key difficulties when annotating rhetorical moves: a decision problem and a consistency problem. The annotation protocol codified in the annotation guidelines aims to provide annotators with guidelines or principles to deal with both difficulties.

The final annotation protocol was developed using guidelines and suggestions in the research literature (Blake, 2018; Biber, Connor, and T. A. Upton, 2007; Garside, Leech, and McEnery, 1997; Hovy and Lavid, 2010; T. Upton and M. Cohen, 2009; Volodina et al., 2014).

Step 6: Annotation guidelines (coding booklet)

Multiple versions of the annotations guidelines were created using both synthetic and holistic approaches (Seliger and Shohamy, 1989, pp.25–29). Synthetic approaches tended to prescribe labels based on particular verbs. In many instances, these prescriptions were accurate, but not on every occasion. If particular strings of characters were indicators of a move, a rule-based parser could be used to automatically label moves. However, it was not possible to create such a rule-based system. The central problem was that words that appear to indicate a particular move were found to also indicate other moves. For example, the presence of the word *method* may indicate a METHOD MOVE, but it could occur in the INTRODUCTION MOVE to describe past research or in the RESULT MOVE when the product is a method. The final version of the annotation guidelines adopts a holistic approach with guiding questions and principles rather than prescribing indicator phrases. The full guidelines are provided in Appendix A.2.

Hovy and Lavid (2010, p.19) point out that:

> Every annotation instantiates some theory. The theory provides the basis for the creation and definition of the annotation categories as well as for the development of the annotation scheme and guidelines, a preliminary and fundamental task in the annotation process. The more complex the phenomena being annotated, the more complex the theory generally is, and hence the more complex the instructions to the annotators.

Annotating at the functional-semantic level is complex, necessitating extensive guidelines. Detailed annotator guidelines were created following the recommendation of Leech and Eyes (1997, p.38) and the procedure used by Volodina et al. (2014). The importance of such guidelines is affirmed by Fort, Nazarenko, and Rosset (2012, p.897), who asserts that the "annotation guide is recognized as the keystone of annotation campaigns".

The final annotation guideline booklet comprises nine pages with over 3200 words. The annotation procedure, nomenclature and annotation schema are introduced. For each move there is a summary table, indicative examples, a description of possible problems, a list of potential indicators of the presence of the move and a discussion of boundary cases.

Step 7: UAM CorpusTool guide

Based on a pilot study with double annotators, it was found that the UAM CorpusTool was not user-friendly. A tailor-made guide was created to simplify the annotation task by providing detailed guidance for the steps used in this process. Screenshots are provided to make the guide easier to follow. The complete guide to using the UAM CorpusTool is provided in Appendix A.20.

4.4.3 Annotation stage

Annotation steps

The first step in the annotation stage was to annotate. The second step was to verify the quality and completeness of the annotations using a tailor-made online tool, the *Move Highlighter*. The third step was to consult specialist informants either face-to-face or online and seek their input on the veracity of the annotations. After discussion, any changes to the annotations were made immediately.

Step 1: Annotation

Each abstract was annotated at least once by the primary researcher. Abstracts that were rather opaque due to terminology overload were checked multiple times over this study. With the added familiarity of the genre, and a better grasp of the disciplinary knowledge, abstracts annotated in the early stages of the study were revised.

Revisions to the initial annotations were made for three reasons, namely:

1. improved subject knowledge casting new light on the content;
2. feedback from specialist informants; and
3. discovery of classification decisions that deviated from the SOP detailed in the annotation guide.

An example of an annotated abstract is given in Figure 4.8.

Step 2: Verification by annotator

After annotating each set of abstracts, the annotator performed a self verification to check that the annotations were complete and accurate. Given the difficulty of reading annotated texts within the UAM CorpusTool, an online visualization tool called the *Move Highlighter* was created. A screenshot of the *Move Highlighter* is shown in Figure 4.9.

The *Move Highlighter* uses regular expressions to match the annotation tags embedded in XML files. On matching, the name of the move and sub-move (if any) are displayed as labels prior to each sentence. By coding at one level finer, namely sub-move, this is envisaged to increase the accuracy of the next level of granularity, i.e. move. To enhance readability, the labels were colour coded using a rainbow-inspired theme as shown in Figure 4.10 so that annotators can easily notice rhetorical moves that might have been accidentally coded incorrectly as the colour order would not follow those in the visible spectrum.

```
<?xml version='1.0' encoding='utf-8'?>
<document>
<header>
<textfile>Trans on Image Processing/Tr_on_ImageProcess_1.txt</textfile>
<lang>english</lang>
</header>
<body>
<segment features='problem;introduction;rhetorical_moves' state='active'>We address the problem of model-based object recognition.</segment> <segment features='purpose;rhetorical_moves' state='active'>Our aim is to localize and recognize road vehicles from monocular images or videos in calibrated traffic scenes.</segment> <segment features='method;rhetorical_moves' state='active'>A 3-D deformable vehicle model with 12 shape parameters is set up as prior information, and its pose is determined by three parameters, which are its position on the ground plane and its orientation about the vertical axis under ground-plane constraints.</segment> <segment features='purpose;rhetorical_moves' state='active'>An efficient local gradient-based method is proposed to evaluate the fitness between the projection of the vehicle model and image data, which is combined into a novel evolutionary computing framework to estimate the 12 shape parameters and three pose parameters by iterative evolution.</segment> <segment features='background;introduction;rhetorical_moves' state='active'>The recovery of pose parameters achieves vehicle localization, whereas the shape parameters are used for vehicle recognition.</segment> <segment features='method;rhetorical_moves' state='active'>Numerous experiments are conducted in this paper to demonstrate the performance of our approach.</segment> <segment features='methodasproduct;result;rhetorical_moves' state='active'>It is shown that the local gradient-based method can evaluate accurately and efficiently the fitness between the projection of the vehicle model and the image data.</segment> <segment features='resultasproduct;result;rhetorical_moves' state='active'>The evolutionary computing framework is effective for vehicles of different types and poses is robust to all kinds of occlusion.</segment> </body>
</document>
```

FIGURE 4.8: Annotated abstract from Image Processing [IP 001]

Step 3: Verification by specialist informants

The UAM CorpusTool is not very user-friendly and so to reduce the burden for specialist informants, a tailor-made online visualization tool was created: the *Move Visualizer*. This tool works in any browser and so enables specialist informants to visualize the full set of annotated abstracts within their discipline without needing to install any software. The XML files of each abstract were housed in cloud storage and populate the interface on demand. In the same way as the *Move Highlighter*,


 It is shown that the local gradient-based method can evaluate accurately and efficiently the fitness between the projection of the vehicle model and the image data. The evolutionary computing framework is effective for vehicles of different types and poses is robust to all kinds of occlusion. </body>
</document>
Highlight

FIGURE 4.9: *Move Highlighter* Interface

FIGURE 4.10: Rainbow colour scheme for rhetorical moves

colour-coded rainbow-inspired labels are used for each move. Figure 4.11 shows the *Move Visualizer* with an abstract from image processing loaded from cloud storage. Specialist informants can click the "show annotations" toggle buttons to execute a script that uses regular expressions to match and colorize the annotations as shown in Figure 4.12.

When specialist informants want to suggest changes and comment on the annotations, they leave comments using the online interface (see Figure 4.13). The comments are stored securely. All suggestions are considered carefully to increase the accuracy of the final dataset.

4.4.4 Double annotation stage

Double annotators were recruited, trained, benchmarked and supported throughout the annotation process. After the annotations were completed, inter-annotator agreements measures were calculated.

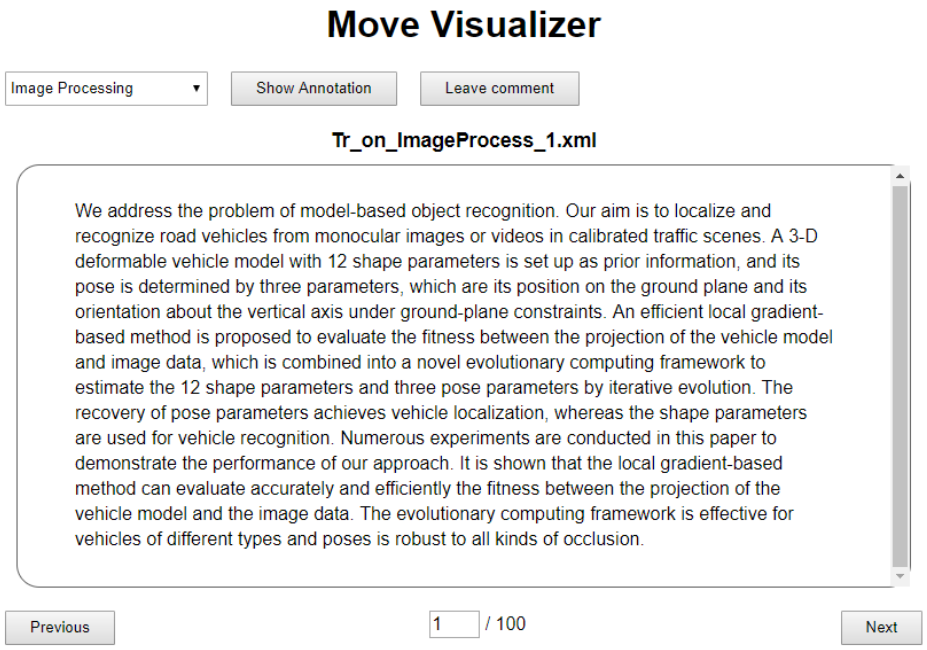


FIGURE 4.11: *Move Visualizer* with image processing abstract selected

Step 1: Recruitment

Based on the results of the pilot double annotation, annotators who had subject knowledge but did not hold senior faculty positions were recruited as double annotators. Two categories of annotators were recruited: (1) enthusiastic Ph.D. candidates, and (2) linguists who also held degrees in science. Double annotators with disciplinary knowledge were used where possible. Seliger and Shohamy (1989, p.52) notes the necessity for researchers to have prerequisite background knowledge, which in this research, based on extensive trials, was deemed to be disciplinary knowledge. All annotators were holders of master's degrees and most were PhD candidates or post-doctoral researchers.

Step 2: Training course and benchmarking

In addition to the final version of the coding booklet, the double annotators were given the UAM CorpusTool Guide. An online training course was created to help double annotators better understand the annotation guidelines and provide annotators with additional practice. This course was initially uploaded on Moodle, a learning management system (LMS). Feedback on the training course housed on the Moodle LMS was received from two colleagues prior to release. The course was later rehoused on Schoology, a different LMS due to institutional constraints. Figure 4.14 shows a screenshot of the Schoology[1] version of the course.

In the online training course, achievement tests were provided at the level of move and a final benchmark proficiency test was also included. In order to qualify

[1]https://www.schoology.com/

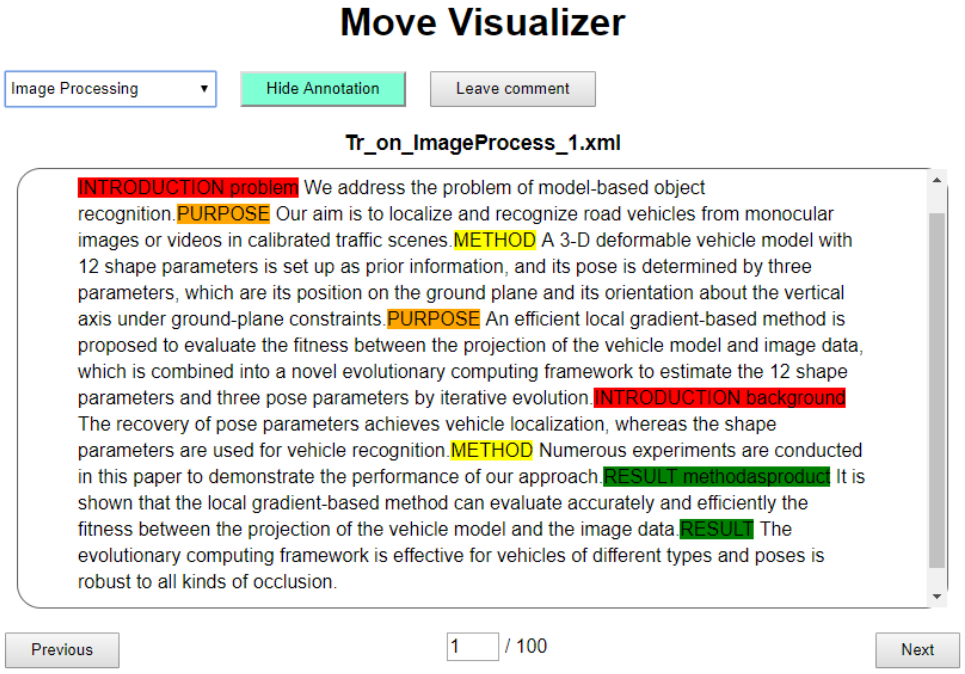


FIGURE 4.12: *Move Visualizer* with annotations visualized

to annotate, annotators were expected to achieve 85% or higher on the benchmark test. Potential annotators were allowed to retrain and retake the benchmark once if their score was between 70% and 84.9%. However, those who failed to achieve 70% on the first benchmark were not allocated any annotation tasks and not invited to retry the benchmark test. Candidate annotators who completed the online training and passed the benchmark test were selected as double annotators.

Step 3: Double annotation

Each of the specialist annotators labelled moves in their own discipline. In total, 10% of the corpus was double annotated, which is in line with similar studies (Hyland, 2004).

Step 4: Inter-annotator agreement

Measures of inter-annotator agreement are used to monitor the reliability of the annotators. The measures quantify not only agreement but disagreement. These measures can be used for both intra- and inter-annotator agreement although in this research only inter-annotator agreement is measured. Artstein and Poesio (2008) asserts that reliability is a prerequisite to establish the validity of any coding scheme", but reliability itself does not imply validity. Unreliability, however, can rule out validity.

Comparing the observed agreement alone does not take into account chance agreement. Chance agreement is the expected percentage of agreement that would happen

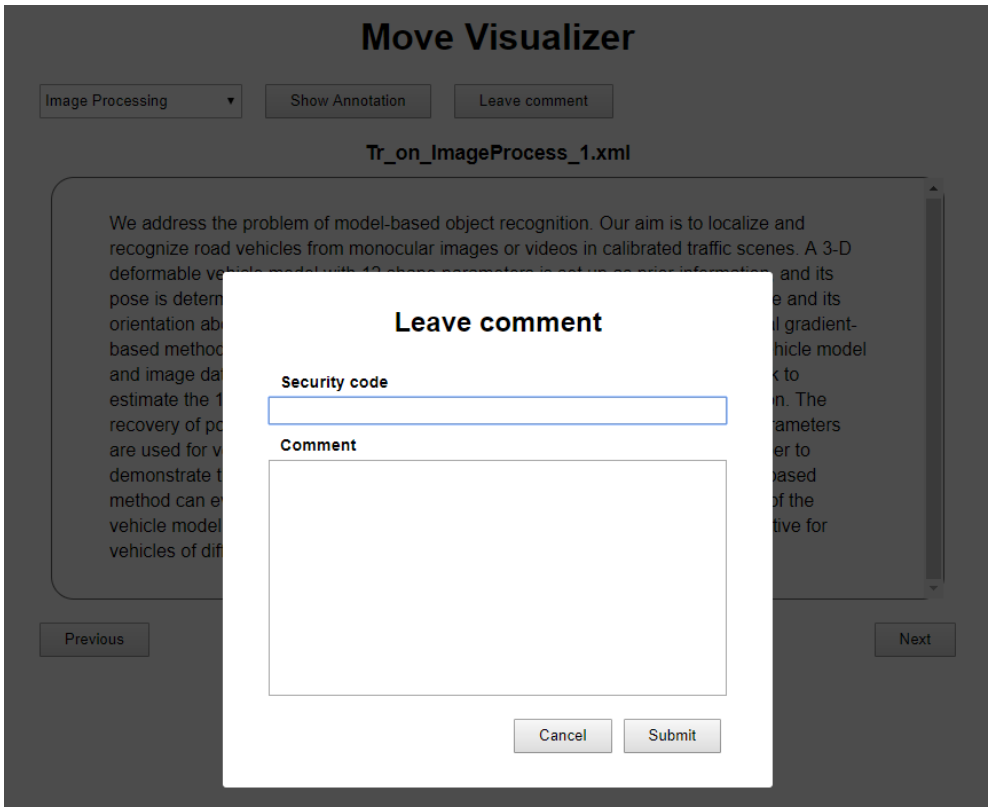


FIGURE 4.13: Comment function in *Move Visualizer*

by chance, and so to correct for this a statistical measure can be harnessed. One such measure is Cohen's kappa coefficient (J. Cohen, 1960). Complete agreement between annotators is shown when the value of κ is 1. Conversely, complete disagreement is shown when the value of κ is 0. The Cohen's kappa measure is suitable when comparing two annotators. In this study, annotations for 10% of the corpus were compared. A Cohen's kappa coefficient of 0.835 was achieved. According to (Landis and Koch, 1977), coefficient scores between 0.81 and 1 may be characterized as almost perfect agreement. Having obtained the statistical measure, the next step is to interpret its meaning. However, Mathet et al. (2012, p.810) report that interpretation of of IAA reliability scores is "highly arbitrary" with "Kappa coefficients [being] difficult to compare, even within the same annotation task". A kappa coefficient score of over 0.81 is considered as "almost perfect agreement".

4.5 Analysis phase

4.5.1 Overview

The analysis phase comprises of a number of types of analysis. The main analyses are listed below.

1. statistical analysis of the corpus
2. *comparative analysis* of the organization of rhetorical moves

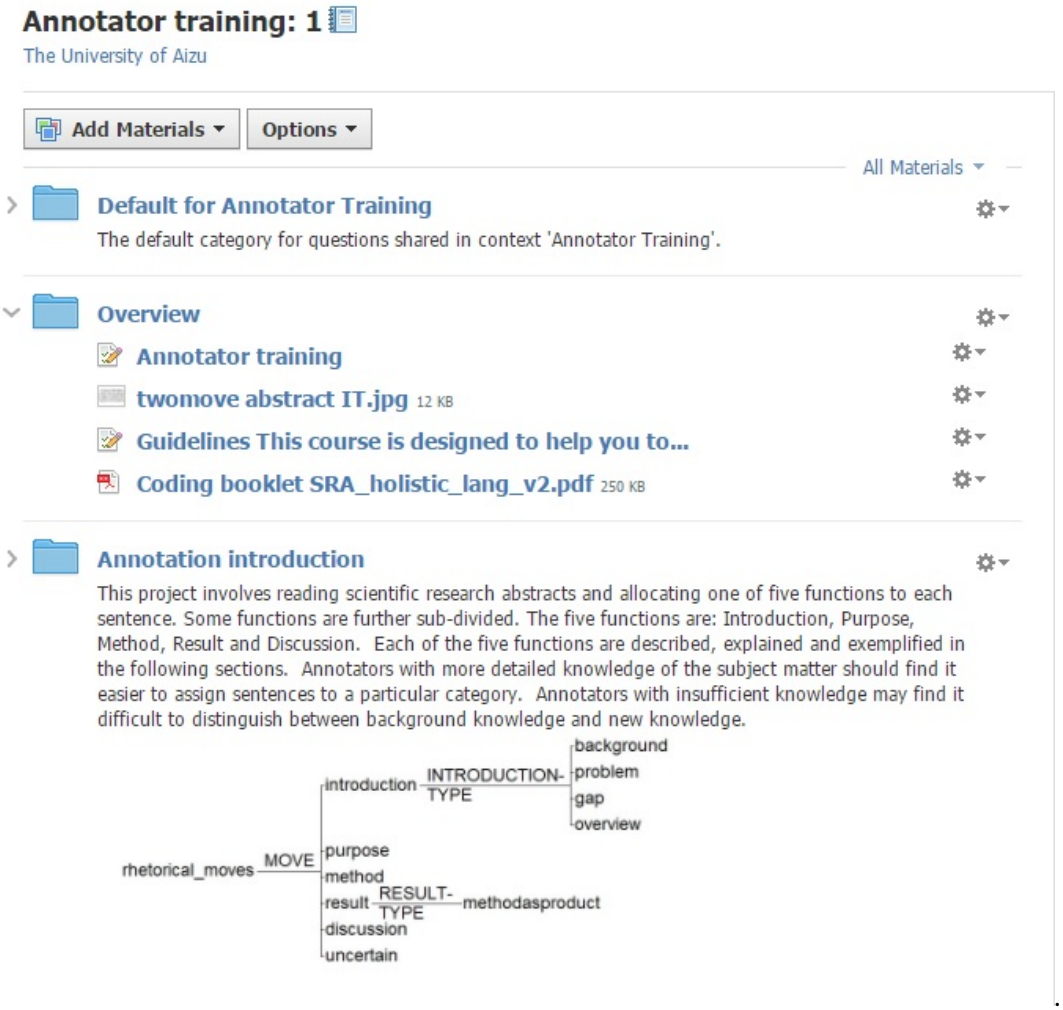


FIGURE 4.14: Screenshot of administrator view of online annotator training course

3. *multidimensional scaling* and cluster analysis *of permutations of rhetorical moves*
4. *key word analysis* of lexis within rhetorical moves
5. *tense analysis* within rhetorical moves
6. *multidimensional scaling and cluster analysis* for rhetorical moves and lexical realization

Each of the main research questions is divided into sub- research questions, which simplify the operationalization of the research. To answer the sub-research questions that use hypothesis-testing, hypothetico-deductive reasoning was adopted in which the null hypothesis is tested to determine its acceptance or rejection. To address sub-research questions that were exploratory, inductive and abductive reasoning were used.

Pre-processing in the analysis phase involved extracting data and manipulating the data into a form that can be used in statistical analysis. In this project, the process was straightforward. The annotations were exported from the UAM Corpus Tool

in XML Extensible Markup Language (XML) which is used to encode documents so both they can be read by both humans and machines. A "tidy data" approach was adopted as advocated by Wickham (2014). The tidy data approach has three components, namely:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

4.5.2 Functions in R for comparative analysis

R for statistics was used as the primary statistical environment. The aims of the R scripts for rhetorical organization are threefold:

1. to count the frequency of moves;
2. to count the number and frequency of permutations of moves; and
3. to compare the differences among and between the datasets.

Following the advice of Gries (2013, p.58), the R scripts that were specially-written for this project were initially drafted in integrated development environments with syntax highlighting enabled, and then were transferred to the R programming environment. Commands in R comprise two elements, namely a *function* followed by *arguments* in parentheses. The arguments specify how and to what the function is applied (ibid., p.61). Multiple functions were created. The functions are designed so that when arranged in particular sequences, scripts that address the research aims are created. Using functions makes the code easier to maintain by reducing repetition. An overview of the main functions is given below.

Function 1: Create master dataframe

- Import annotations (features) and sentences (segments) from all texts (1000 XML files stored in one folder named AllTexts) into RawMaster dataframe.

Function 2: Manipulating data

- Remove unnecessary columns.
- Split feature column into two columns (feature_move and feature_sub-move).
- Simplify terms from all rows in RawMaster (e.g. Tr_on_ImageProcess –> IP)
- Save as Master dataframe.

Function 3: Create feature only dataframe

- Remove segments column from Master and save as Feature dataframe

Function 4: Create feature specific dataframes

- Extract all rows of texts annotated with particular moves from Master dataframe
- Save as Feature_Move dataframe.
- Extract all rows of texts annotated with particular sub-moves from Master dataframe
- Save as Feature_sub-move dataframe.

Function 5: Create discipline specific dataframes

- Extract all rows of texts (n = 100) of each discipline from Master into separate data frames
- Name each row according to discipline Name_Discipline (e.g. IP_Discipline, WC_Discipline, KDE_Discipline)

Function 6: Increment count of feature instances

- Increment count of occurrences of features in (1) each discipline (Name_Discipline) dataframe
- Increment count of occurrences of features in Master dataframes.

Function 7: Extract raw feature permutations

- Extract sequence of all feature_move or feature_sub-move for each text to create and create new dataframe. (e.g. for Abstract IP_n, with 5 sentences introduction, background; method; method; results, methodasresult)
- name dataframe as MasterPermutation dataframe.

Function 8: Abbreviate feature permutations

- Replace the names of the features with capital letters (e.g introduction = I, method = M) from MasterPermutation Dataframe
- Replace the names of the sub-moves to lowercase letters (e.g. gap = g, background = b) from MasterPermutation Dataframe
- Store results in new AbbreviatedPermutation dataframe

Function 9: Merge sequential identical permutations

- In the AbbreviatedPermutation dataframe, when two identical features occur sequentially, omit the second feature. (e.g. for Abstract IP_n {Ib, M, M, Rm} becomes {Ib, M, Rm}
- save as MergedPermutation dataframe.

Function 10: Omit sub-moves in permutations

- In the AbbreviatedPermutation dataframe omit all sub-moves.(e.g. for Abstract IP_n {I, M, M, R}
- Save as FiveMovePermutation dataframe.

Function 11: Omit sub-moves in permutations

- In the FiveMovePermutation dataframe convert P to I
- Run merge script 9
- Save as FourMovePermutation Dataframe.

Function 12: Increment count of permutation instances

- Increment count of occurrences of permutation in all ten disciplines in AbbreviatedPermutation dataframe
- Increment count of occurrences of permutation in all ten disciplines in MovePermutation dataframe.

Function 13: Increment linearity count

- Set expected sequence of moves.
- Features occurring in the expected order are assigned a value of *O*.
- Features occurring in the expected order are assigned a value of 1.
- Run on the Merged Permutation, FiveMovePermutation and FourMovePermutation dataframes.

The expected sequence of moves is set as: Introduction, Purpose, Method, Result and then Discussion. The temporary category of Uncertain is ignored. There is no expected sequence of sub-moves.

Function 14: Increment cyclicity count

- Identify reoccurrence of move sequences.
- Increment the cyclicity count by 1
- Run on the Merged Permutation, FiveMovePermutation and FourMovePermutation dataframes.

4.5.3 Basic R scripts for comparative analysis

The functions detailed in the previous subsection can be sequenced to realize the functionality required. The following is the plan of how the functions are stacked to create scripts. The R scripts are available in Appendices A.14, A.16, A.17, and A.18.

Script 1: Feature frequency

- Function 1: Create RawMaster dataframe (import)
- Function 2: Create Master dataframe (manipulate)
- Function 6: Count feature instances in Master dataframe by discipline

Script 2: Comparison and contrast of text feature

- Function 1: Create RawMaster dataframe (import)
- Function 2: Create Master dataframe (manipulate)
- Function 4: Create Feature_Move and Feature_sub-move dataframes (divide)

Script 3: Sequence frequency

- Function 1: Create RawMaster dataframe (import)
- Function 2: Create Master dataframe (manipulate)
- Function 7: Create MasterPermutation dataframe (collect sequences)
- Function 8: Create AbbreviatedPermutation dataframe (abbreviate)
- Function 9: Create MergedPermutation dataframe (merge)
- Function 10: Create FiveMovePermutation dataframe (omit sub-moves)
- Function 11: Create FourMovePermutation dataframe (omit sub-moves)
- Function 12: Count permutation instances in MergedPermutation, FiveMovePermutation and FourMovePermutation dataframes by discipline.

Script 4: Comparison and contrast of sequences

- Function 1: Create RawMaster dataframe (import)
- Function 2: Create Master dataframe (manipulate)
- Function 7: Create MasterPermutation dataframe (collect sequences)
- Function 8: Create AbbreviatedPermutation dataframe (abbreviate)
- Function 9: Create MergedPermutation dataframe (merge)

- Function 12: Count permutation instances in MergedPermutation, FiveMovePermutation and FourMovePermutation dataframes by discipline. (calculate variation index)
- Function 13: Count linearity instances (calculate linearity index)
- Function 14: Count cyclicity instances (calculate cyclicity index)

4.5.4 Multidimensional scaling and cluster analysis

Multidimensional scaling (MDS) and cluster analysis family of methods are used for exploratory analysis rather than hypothesis-testing (Gries, 2013, p.336).

Multidimensional scaling is a way to reduce dimensionality in a non-linear manner. This enables multidimensional data (i.e. data with three or more dimensions) to be mapped into a two-dimensional plane (Cartesian space). This involves computing a distance matrix and visualizing that in the Cartesian space.

The *K*-means clustering method of unsupervised machine learning was selected to identify clusters among the data objects. In *k*-means clustering the number of clusters is set prior to the clustering while in hierarchical clustering the number of clusters is not predetermined. *K*-means clustering is the most established and easy-to-use method that can be implemented with relatively straightforward code using Python. Algorithm 1 shows how *K*-means clustering is operationalized using easy-to-read pseudocode. Clusters are created by calculating the geometric centre, called the centroid, between and among data points. The process begins with just two data points but continues until all data points are considered. As each data point is added to a cluster, the centroids for all clusters are recalculated. This process continues until there are no changes in the position of the centroids, and the cluster groupings are therefore finalized.

Algorithm 1 Basic *K*-means clustering algorithm using random seed

1: Set the number *k* of clusters to assign;
2: Initialize *k* centroids using random seed;
3: Assign data point to closest centroid;
4: Compute new centroid for each cluster;
5: Repeat steps 3 and 4 until centroid positions do not change.

The script used for MDS and cluster analysis is available in Appendix A.8.

4.5.5 Keyness and key word analysis

The primary method to investigate Keyness was through the *Keyness calculator* while the secondary method was via the keyness function built into AntConc (Anthony, 2019). (See Subsection 2.6.4 for a more detailed description of keyness.)

For the multidimensional analysis and cluster analysis, Keyness was calculated using the Keyness calculator, an open-source Python script (See Appendix A.9 for the script).

A particular strength of this script is that it allows the use of the simple math formula Ratio (Kilgarriff, 2009). The reference corpus for each sub-corpus was set to the whole corpus.

AntConc (Anthony, 2019) was used to investigate the annotated corpus for numerous standard corpus queries, such as examining words in context, plotting dispersion, arranging words by frequency and ranking key words by keyness. The reference corpus using for AntConc was the Brown corpus and the association measure selected was the default Log likelihood score.

4.5.6 Tense analysis

The extant literature on the identification of grammatical tenses is sparse. Even in computational linguistics, this is a subfield that has attracted little attention. TimeML (Pustejovsky et al., 2003) is the most well-known system to label tense, but does not discriminate between the twelve forms. Yule (1998) notes textbooks dedicated to teaching English to non-native speakers focus on the twelve verb forms, or grammatical tenses. In the grammatical tense system of English future, progressive forms and perfect forms are not considered as tense, but as aspect. In this sense, tenses are objective while aspect is subjective and changes according to stance adopted. Georgiev (2006) created an algorithm to identify tense (to an unknown degree of accuracy) for the propriety software Syntparse (Lumsden, 1994) but this software is no longer available.

It was, therefore, necessary to create a tailor-made program to identify tense. A program in Python that utilises the tokenisation and part-of-speech (POS) tagging using the Natural Language Toolkit (NLTK) (Bird, 2006; Bird, Loper, and Klein, 2009; Loper and Bird, 2002) was created (See Appendix A.11). The verb group can be identified by the POS tags, and a decision tree, key-value dictionary or regular expressions can then used to match particular permutations of parts of speech (POS) and word forms to identify the grammatical tense of the verb group. The tagset used is the Penn Treebank tag set (Marcus, Santorini, and Marcinkiewicz, 1993). The full set of tags are given in Figure 4.15.

By analysing the structure of finite verb groups in declarative statements, a matrix was created that shows the presence or absence of particular parts of speech by grammatical tense. This matrix is shown in Table 4.3. Although there are twelve tenses, there are 18 combinations of POS tags that can form these grammatical tenses.

From the table of combinations of part-of-speech (POS) tags, a parse tree can be created that specifies the permutations of POS tags that result in a particular tense. Figure 4.16 shows an extract of the parse tree that was created.

The complete parse tree can be found in Appendix A.4. The parse tree was used as the basis to create a decision tree algorithm in Python. Decision trees start with a node that represents a feature or attribute, which in this case is a part of speech (POS), the branch represents a decision rule (e.g. the presence or absence of a POS tag), and each leaf node represents the outcome (e.g. the tense identified).

- CC coordinating conjunction
- CD cardinal digit
- DT determiner
- EX existential there
- FW foreign word
- IN preposition/subordinating conjunction
- JJ adjective 'big'
- JJR adjective, comparative
- JJS adjective, superlative
- LS list marker
- MD modal
- NN noun, singular
- NNS noun plural
- NNP proper noun, singular
- NNPS proper noun, plural
- PDT predeterminer
- POS possessive ending
- PRP personal pronoun
- PRP$ possessive pronoun
- RB adverb
- RBR adverb, comparative
- RBS adverb, superlative
- RP particle
- TO infinite
- UH interjection
- VB verb, base form
- VBD verb, past tense
- VBG verb, gerund/present participle
- VBN verb, past participle
- VBP verb, sing. present, non-3d
- VBZ verb, 3rd person sing. present
- WDT wh-determiner
- WP wh-pronoun
- WP$ possessive wh-pronoun
- WRB wh-abverb

FIGURE 4.15: Penn Treebank tag set

The code to discriminate between the twelve grammatical tenses is provided in Appendix A.11. However, to provide some insight into how the code works an extract of the code is given in Figure 4.17. In this block of code, we can focus on the node part of speech *VBP*, single present non-third person verb [line 9]. Once this POS tag is identified as node, the code searches for the POS tags for the subsequent words. The sequence of POS tags are given in lines 17 to 20. There are four children in the code block. Two are passive voice and two are active voice.

The first child is used to match the sequence of *VBP* followed by *VBN* (past participle) [line 17]. The only tense realized by this sequence of two POS tags is *present perfect simple*, and so that tense is then assigned to that sequence. However, if the following tag is *VBN*, the tense is changed to *present perfect simple* in passive voice [line 18] while if the following POS tag is *VBG*, the tense is changed to *present perfect progressive* [line 19]. Finally if the four POS tag sequence of *VBP, VBN, VBG* and *VBN* is matched, the tense is assigned as *present perfect progressive* in passive voice [line 20].

For the complete script please refer to A.11.

4.6 Chapter Summary

Based on a thorough review of the literature on research methodology and corpus linguistics in particular, the research methodology was designed. A three-phase procedure was decided upon comprising corpus, annotation and analysis phases.

In the corpus phase, ten disciplines were carefully selected. The disciplines were selected to create a balanced corpus with an emphasis on the under-represented hard-to-read disciplines, such engineering and applied sciences. The criteria for selection of disciplines, journals and texts were developed. Protocols for operationalization such as standard procedures for corpus collection and cleaning were documented.

TABLE 4.3: Matrix of part of speech tags for the twelve grammatical tenses

Tense	Form of tense	VBP	VBZ	VB	VBG	VBN	VBD	MD
Present simple	non-3rd person verb	1	0	0	0	0	0	0
Present simple	3rd person verb	0	1	0	0	0	0	0
Present simple	be (used after modal MD)	0	0	1	0	0	0	0
Present progressive	(am | are)+ being	1	0	0	1	0	0	0
Present progressive	is + being	0	1	0	1	0	0	0
Present progressive	be + being	0	0	1	1	0	0	0
Present perfect simple	have + been	1	0	0	0	1	0	0
Present perfect simple	has + been	0	1	0	0	1	0	0
Present perfect progressive	(has | have) + been + Ving	1	0	0	1	1	0	0
Present perfect progressive	(has | have) + been + Ving	0	1	0	1	1	0	0
Past simple	past form	0	0	0	0	0	1	0
Past progressive	(was | were)+Ving	0	0	0	1	0	1	0
Past perfect simple	had + Vpp	0	0	0	0	1	1	0
Past perfect progressive	had + been + Ving	0	0	0	1	1	1	0
Future simple	will + base form	0	0	1	0	0	0	1
Future progressive	will + be + Ving	0	0	1	1	0	0	1
Future perfect simple	will + have + Vpp	0	0	1	0	1	0	1
Future perfect progressive	will + have + been + Ving	0	0	1	1	1	0	1

The annotation phase includes descriptions of the annotation protocol including the detailed annotation guide specifically created for this project, and the development of a training procedure for double annotators, and measures to evaluate inter-annotator reliability. Manual annotation of rhetorical moves was conducted by the principal investigator. Double annotators were recruited, trained and benchmarked. Inter-annotator agreement measures provide a window through which the veracity of the annotations can be estimated.

In the analysis phase, software programs were developed to extract and analyze the annotated labels and entities. As this was a multi-year project, the programs were continuously refined to increase their accuracy. Programs were created using C, Visual Basic for Applications, R and Python. When available, ready-to-use functions were deployed; but in order to directly address the research questions, multiple tailor-made scripts were created. The lexical realization was investigated using keyness as a proxy of lexis and grammatical tense as a proxy for grammar. The similarity and differences in lexical realization were analyzed using *k*-means hierarchical clustering for tense, for keyness, and for both tense and keyness together.

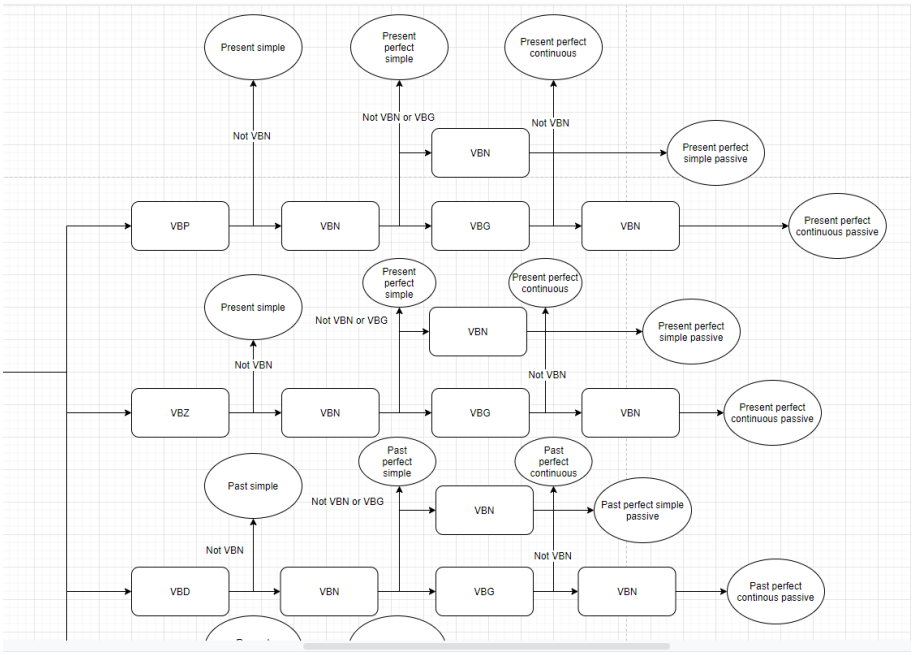


FIGURE 4.16: Extract of parse tree diagram

```
from anytree import Node, RenderTree
from anytree.search import findall_by_attr

import copy

root = Node("root")
#root children

vbp = Node("vbp", parent=root, tense="pressimp")
vbz = Node("vbz", parent=root, tense="pressimp")
vbd = Node("vbd", parent=root, tense="pastsimp")
tobePres = Node("tobepres", parent=root, tense="pressimp")
tobePast = Node("tobepast", parent=root, tense="pastsimp")
md = Node("md", parent=root, tense=None)

#vbp children
vbp_vbn = Node("vbn", parent=vbp, tense="presperfsimp")
vbp_vbn_vbn = Node("vbn", parent=vbp_vbn, tense="presperfsimppass")
vbp_vbn_vbg = Node("vbg", parent=vbp_vbn, tense="presperfcont")
vbp_vbn_vbg_vbn = Node("vbn", parent=vbp_vbn_vbg, tense="presperfcontpass")
```

FIGURE 4.17: Extract of tense identification script

Chapter 5

Rhetorical organization

> *There are only patterns, patterns on top of patterns, patterns that affect other patterns. Patterns hidden by patterns. Patterns within patterns.*
> - Chuck Palahniuk, American Journalist and author of Survivor

5.1 Chapter preview

This chapter describes and discusses the results that relate to the first research question (2.8.2), namely: "What is the rhetorical organization of abstracts of research articles published in a broad range of top-tier scientific journals?"

This research question is divided into six sub-questions, which are reproduced here:

1. What moves occur in research abstracts in each discipline?
2. How frequent is each move in research abstracts in each discipline?
3. In what sequence do the moves occur in research abstracts in each discipline?
4. How frequent is each sequence in research abstracts in each discipline?
5. What are the similarities in rhetorical organization between the disciplines?
6. What are the differences in rhetorical organization between the disciplines?

Section 5.2 provides an overview of the corpus dimensions for the tailor-made corpus of scientific research abstracts. Descriptive statistics are presented for word tokens, word types, sentence length and readability. This statistical analysis of the corpus provides the context to situate the findings discussed in the remainder of this chapter. The following six sections describe and discuss the results related to each of the six sub-questions. Section 5.3 describes the types of moves that were identified in research abstracts for each scientific discipline. Section 5.4 analyzes the frequency of the moves discovered in the corpus of abstracts for each scientific discipline. Section 5.5 extends the analysis of rhetorical moves by investigating the sequences in which rhetorical moves occur in each of the ten disciplines. Section 5.6 considers the frequency of the different permutations within each discipline and

among the disciplines. Section 5.7 focuses on the similarities found in the types of move and their respective frequencies, and the particular move permutations and their respective frequencies. Similarity is investigated using multidimensional scaling and cluster analysis. Section 5.8, in contrast, focuses of the differences between and among the disciplines in terms of moves and permutations of moves. A novel framework using the three-interlocking (Borromean) rings as the vector space to map research abstracts is presented. The three variables of linearity, cyclicity and variation can be mapped to identify the degree of similarity or dissimilarity between disciplinary corpora. This section extends the discussion of the implications and applications of the results on rhetorical organization are enumerated in section 5.9. The chapter summary 5.10 aims to bring together the most important findings and insights made on those discoveries.

5.1.1 Tool selection

> *If the only tool you have is a hammer, you tend to see every problem as a nail.*
> - Abraham Maslow, American psychologist

The *law of the instrument* (Kaplan, 2017, p.28) states that jobs tend to be adapted to the tools rather than adapting tools to the job at hand. However, for this study, the most appropriate tools were selected and where necessary tailor-made tools were created. This subsection provides an overview of the tools selected to investigate rhetorical organization. The purpose of each tool and some pertinent details are provided in Table 5.1.

TABLE 5.1: Tools utilized to investigate rhetorical organization

Purpose	Function within tool	Tool
To count word type and tokens	Word list	AntConc[a]
To count number of sentences	Examine master dataframe	R script[b]
To view 5 number summary	Box plot	R script[b]
To measure readability	Readability function	Language Feature Detector[c]
To check binary values	Presence-absence matrix	Truth table
To visually contrast frequency	Heat map	R script[b]
To visualize categories	Venn diagram	One diagram from set theory
To group similar items	Cluster analysis[d]	R script[b]
To visualize clusters	Dendrogram plot	R script[b]

[a] AntConc (Anthony, 2019) is one of the standard tools to investigate corpora with over 2100 citations on Google scholar
[b] R for Statistics (R Core Team, 2017; Gries, 2013) is a popular statistical environment
[c] The Language Feature Detector is available online from grammar-lexis-viz.herokuapp.com
[d] *K*-means flat clustering algorithm that minimizes Euclidean distance

The word token and word type count values differ depending on the tool chosen due to the differences in the way that each software program tokenizes and categorizes (Anthony, 2013; Blake, 2016). Nevertheless, the approximate proportion between word type and word token will remain the same, and when comparing word type

and word token counts between corpora any differences in counting protocols are highly unlikely to affect any comparison.

The total number of sentences was counted in R (Gries, 2013; R Core Team, 2017) by running the script to create the master dataframe and identifying the row number of the final annotated sentence.

The *Language Feature Detector*[1] is a tailor-made software program which incorporates a readability function. Figure 5.1 shows the results generated for an abstract from Information Theory [IT 25].

The presence-absence matrix is a form a of truth table, which uses Boolean algebra to evaluate the truth function of binary variables. Apart from the visual clarity of seeing zeros and ones, such truth tables make it easy to design software programs to check the value for each cell, and automate the process.

Heat maps were chosen as a data visualization tool to illustrate the comparative frequency and infrequency of particular permutations of adjacent pairs of rhetorical moves. With the advent of easy-to-use data visualization packages, heat maps are becoming more widely used by scientists, (e.g. Bredbenner and Simon, 2019) and linguists (e.g. Dunn, 2019) alike. The statistical environment R incorporates an easy-to-use native heat-map function, which no doubt goes someway to explain the increasing usage of heat maps.

Drawing on ideas from Allwood et al. (1977), set theory can be used to study sets (or groups) of items. Venn diagrams can visualize the data points and/or groups to deduce which items are present in each set or subset.

Multidimensional scaling and k-means cluster analysis are the final tools used. Both multidimensional scaling and cluster analysis are described in depth in Section 4.5.4.

5.2 Corpus dimensions

As the primary aim is to provide a descriptive account of the rhetorical organization of scientific research abstracts, descriptive rather inferential statistics are used. More specifically, one of the central underlying motivations for this study is to be able to provide teachers of scientific research writing and novice scientific writers with accurate disciplinary specific guidance on rhetorical organization. With a more thorough understanding of this important genre of writing, writers of research abstracts should be more able to draft scientific research abstracts that show generic integrity and will be accepted by members of their particular disciplinary community of practice.

This section contextualizes the statistics for frequency of rhetorical moves and permutations of rhetorical move sequences by providing an overview of the corpus dimensions using word token and word type counts, sentence lengths and readability scores. The word token and word type counts help form a concrete schema to understand the scale of the corpus. From sentence length and readability, it is possible

[1] https://grammar-lexis-viz.herokuapp.com/

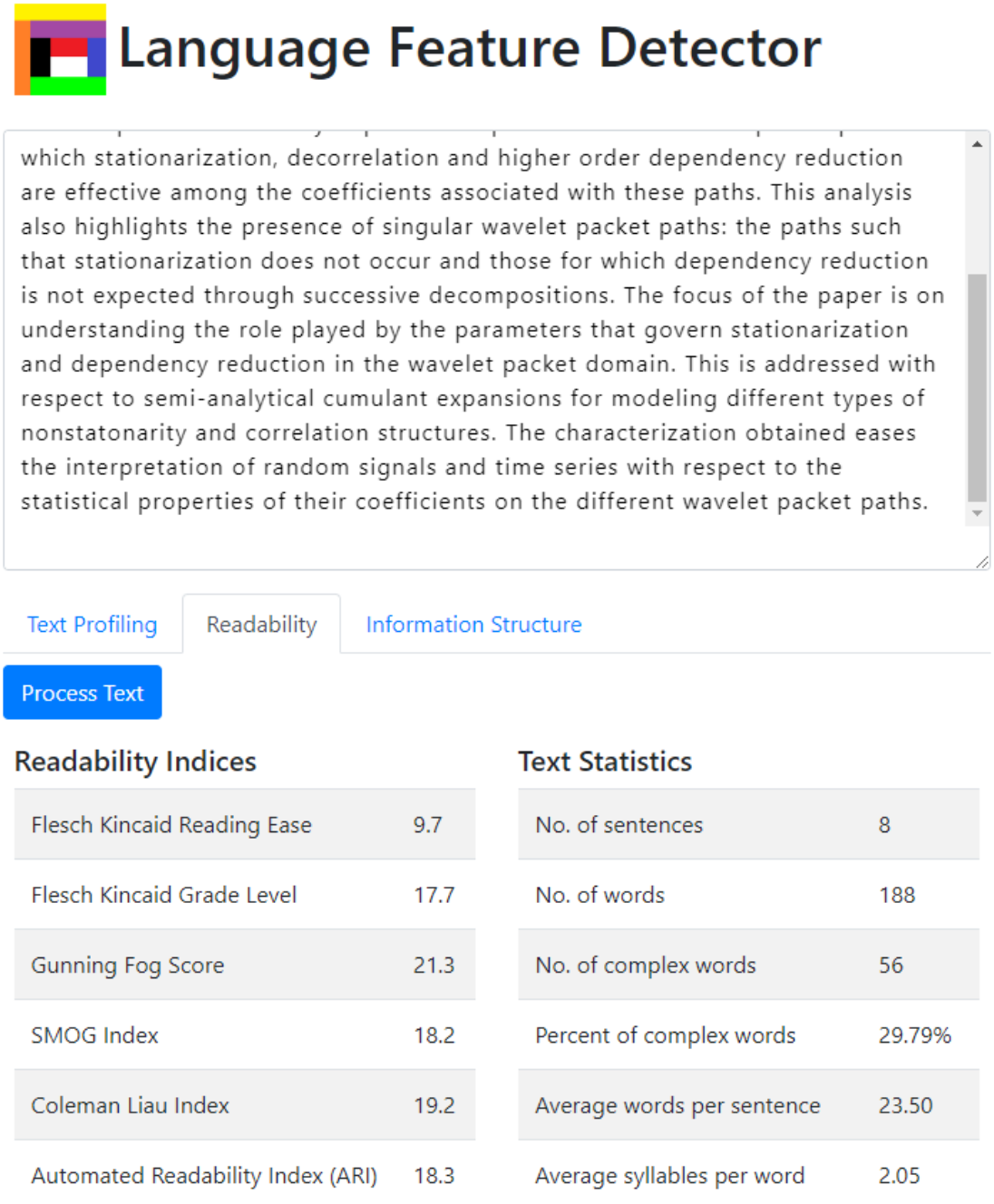


FIGURE 5.1: Output for abstract IT 25 using the readability function of the *Language Feature Detector*

to understand the degree to which the corpus is homogeneous in terms of reading difficulty. The high level of reading difficulty may explain why some of the more technical applied scientific and engineering disciplines, such as IT and WC are largely under-researched.

5.2.1 Word tokens and word types

An isotextual corpus was created with identical numbers of texts per discipline. Table 5.2 provides a numerical overview of the corpus, showing each discipline, the journals from which research abstracts were drawn and the total number of texts (research abstracts) incorporated in the corpus. Each text is a complete written research abstract.

TABLE 5.2: Details of the corpus of scientific research abstracts

Code	Discipline	Journals	Quantity
BOT	Botany	The Plant Cell	100
EC	Evolutionary computing	Transactions on Evolutionary Computing	100
IND	Engineering	Transactions on Industrial Electronics	100
IP	Image processing	Transactions on Image Processing	100
IT	Information theory	Transactions on Information Theory	100
KDE	Knowledge, data and engineering	Transactions on Knowledge, Data and Engineering	100
LING	Linguistics	Journal of Communication, Applied Linguists and Journal of Cognitive Neuroscience[a]	
MAT	Materials science	Advanced Materials	100
MED	Medicine	British Medical Journal	100
WC	Wireless computing	Transactions on Wireless Computing	100

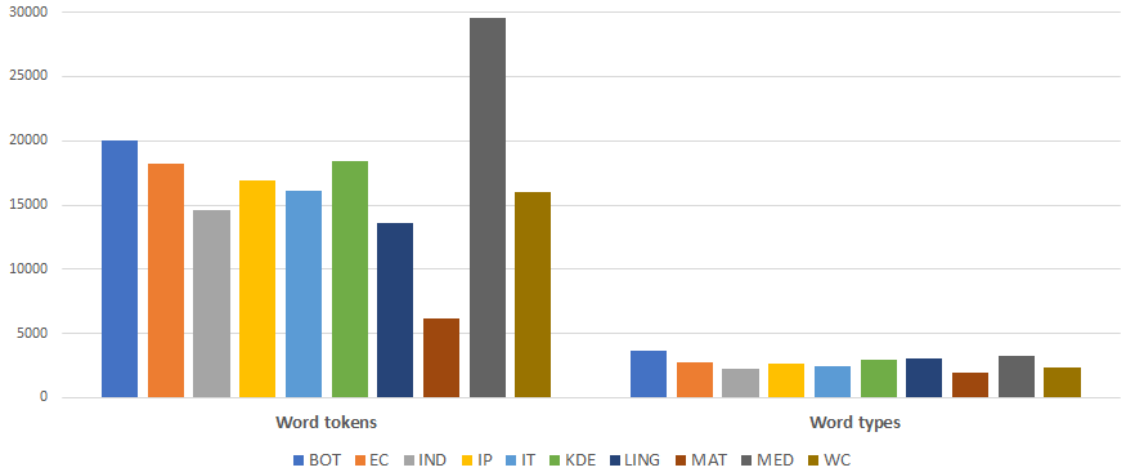
[a] Three journals were selected due to the comparative paucity of research articles in each issue

Figure 5.2 provides a visual overview of the relative sizes in terms of word tokens and word types for each discipline within the corpus. The word token and type counts were carried out using AntConc 3.5.9 (Anthony, 2019). Eight of the disciplines show little variation in the number of word tokens. However, the disciplines of material science and medicine stand out by having substantially different token counts, viz. the lowest and highest number of word tokens respectively.

Table 5.3 provides the numerical counts for word tokens and word types by discipline. The mean number of word tokens per discipline is slightly less than 170,000 while the mean number of word types is 2,723. The mean number of word tokens per abstract is therefore approximately 170. Research abstracts in materials science are substantially shorter in both categories with 6,123 word tokens and 1,947 word types while medical research abstracts have the highest number of word tokens at 29,584 and word types at 3,203.

5.2.2 Sentence length

As described in the methodology, the ontological unit selected for move analysis is the sentence. Table 5.4 shows the number of word tokens, sentences and the mean



BOT = Botany; EC = Evolutionary computing; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; WC = Wireless computing

FIGURE 5.2: Number of word tokens and word types by discipline

TABLE 5.3: Number of word tokens and word types by discipline[a]

Discipline[a]	Word tokens	Word types
BOT	20041	3619
EC	18194	2728
IND	14585	2199
IP	16893	2698
IT	16062	2467
KDE	18443	2947
LING	13602	3035
MAT	6123	1947
MED	29584	3203
WC	15972	2385
Total	169499	12248

[a] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing

TABLE 5.4: Sentence number and length by discipline[a]

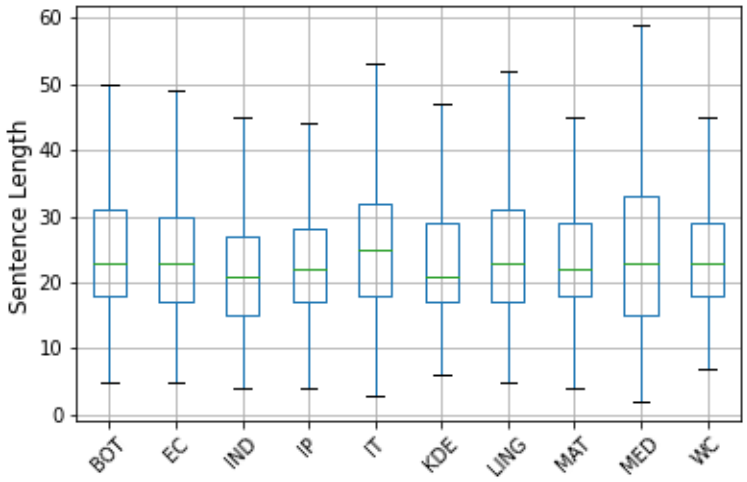
Discipline[a]	Word tokens[b]	Number of sentences[c]	Mean sentence length[d]
BOT	20041	826	24.3
EC	18194	762	23.9
IND	14585	655	22.3
IP	16893	734	23.0
IT	16062	620	25.9
KDE	18443	801	23.0
LING	13602	553	24.6
MAT	6123	262	23.4
MED	29584	1331	22.2
WC	15972	656	23.5
All corpus	169499	7200	23.5[e]

[a] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing
[b] Counted on raw corpus using AntConc 3.5.9 (Anthony, 2019)
[c] Calculated in R on master dataframe based on row number
[d] Reported to one decimal place.
[e] Standard deviation (SD) = 1.08

sentence length by discipline. The mean sentence length varied from 22.2 words in medical abstracts to 25.9 words in information theory abstracts. The mean overall sentence length for all disciplines was 23.5 words. The means for each discipline vary only slightly from the corpus mean, evidenced by a standard deviation (SD) score of 1.08. Sentence length is one of the markers of structural complexity (Biber, 1988) with more elaborate discourses having longer sentences while more restricted discourses have shorter sentences. Given that research abstracts summarize the information in research articles, elaborate and relatively long sentences would be expected to predominate.

Box plots provide a visual summary enabling statistical information to be understood more easily and information conveyed more quickly. The relative positions of the mean and median can be visually compared (Crawley, 2007). Figure 5.3 provides a visual summary of the minimum, maximum and average sentence lengths for each discipline. The horizontal green line across each of the rectangles indicates the mean sentence length for each discipline. There is little fluctuation in the mean value of sentence length, and even the range between the minimum and maximum values are similiar. By a slight margin, medical abstracts show the greatest range from only a couple of words through to sentences with almost 60 words. In the medical abstracts, some moves were realized using sentences that used ellipsis and had no stated finite verb as shown in Example 6, which help explain very short sentences.

(6) Setting UK primary care.
[MED 071]



[a] BOT = Botany; EC = Evolutionary computing; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; WC = Wireless computing
[b] Sentence lengths and averages for the box plot were calculated using a tailor-made Python script (see Appendix A.7)

FIGURE 5.3: Box plot showing average values of sentence length by discipline

TABLE 5.5: Readability by discipline[a]

Discipline	Flesch Kincaid grade level[b]	Flesch Kincaid reading ease[c]
BOT	16.4	15.7
EC	16.5	15.5
IND	16.0	16.8
IP	16.2	16.4
IT	15.4	23.9
KDE	15.2	24.4
LING	16.2	16.3
MAT	17.1	8.7
MED	11.9	32.9
WC	16.9	14.5

[a] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing
[b] This level indicates the grade level (US) of education needed to understand the text. Thus a score of 12, implies grade 12. The higher the score, the less readable the text.
[c] This index estimates the ease of reading on a scale from easily readable (100) to extremely difficult to read (0). Scores below 30 are texts that are very difficult to read.

5.2.3 Readability

As stated in the literature review, research abstracts are informationally and lexically dense, which in turn negatively affects readability. Standard readability statistics, such as the *Flesch Kincaid grade level* (Kincaid et al., 1975) and the *Gunning Fog score* (Gunning, 1969), provide a broad guide to how difficult or easy texts are to read. The *Flesch Kincaid grade level* indicates the grade level that a reader is expected to have reached in order to understand the target text. Thus, low single digit numbers indicate higher degrees of readability than double digit numbers. The *Flesch Kincaid reading ease score* provides a value between zero and 100 with the higher scores representing the easiest to read texts. In short, easy-to-read texts, therefore, receive low scores for *grade level* and high scores for *reading ease*.

Table 5.5 shows the values for these two readability scores namely *Flesch Kincaid grade level* and *Flesch Kincaid reading ease score*. As can be seen in this table, the mean reading grade (based on the US curriculum) for the corpus is 15.8 with a grade range of between 11.9 and 17.1. All the disciplines were classified as needing the reading level of a college graduate, i.e. very difficult to read. The reading ease scores are also in the range of very difficult with a mean score of 18.51 out of 100. Medical abstracts had the highest degrees of readability while materials science abstracts had the lowest. As materials sciences abstracts were the shortest by far, it may be that more nominalization is used to package more informational units, creating very dense text. The medical abstracts were the longest, at around twice the length of the other abstracts, easing the necessity to use multiple prepositional phrases and subordinate clauses. Medical abstracts are aimed not only at researchers but practising medical doctors and so despite the complexity of the technical vocabulary, the texts are designed to be readable. Example 7 highlights the prototypical difficulty reading this dataset: unknown words and concepts.

(7) Conditionally cycle-free graphical models (i.e., cyclic graphical models which become cycle-free after conditioning on a subset of the hidden variables) are constructed for coset codes.
[IT 005]

(8) QD functionalization allows a clear elucidation of pH-tunable drug release and facile visualization of in vivo delivery event.
[MAT 008]

Although it is possible to construct a sentential model of how the concepts interact. Without specialist knowledge, the actual meaning of the sentence is not deducible. The non-specialist reader will simply understand that "some models are constructed for some codes". Example 8 is similar with non-specialist readers decoding the sentence as something explains some type of drug release and visualization of a delivery event. This sentence was shown to five readers, four of whom guessed that it was describing research about births. Yet, the sentence describes visualization

of a drug targetting tumors. Presumably, the presence of of the Latin term *in vivo* and its collocation with *fertilization* combined with the presence of the word *delivery* explained how non-specialists made this incorrect judgement.

5.2.4 Summary

To sum up, this corpus of scientific research abstracts consists of highly technical terminology which are likely to be incomprehensible to lay readers. The corpus is isotextual and relatively homogeneous in terms of sentence length and readability. With the exception of two disciplines, BOT and MED, the length of abstracts are similar. BOT uses graphical abstracts and are considerably shorter while MED uses structured abstracts and are considerably longer.

The following six sections (Section 5.3 to 5.8) discuss the results pertaining to one sub-research question. The first four sub-research questions are hypothesis-testing (Sub-research questions 1, 2, 3 and 4). Sub-research questions 5 and 6 are exploratory and so may be considered as hypothesis-generating.

5.3 Sub-question 1: The types of rhetorical moves

5.3.1 Preamble

This section investigates the first sub-research question using the corpus described in 5.2. Sub-question 1 is:

> "What moves occur in research abstracts in each discipline?"

This question can be reformulated as two hypotheses: the null hypothesis to be tested and an alternative hypothesis to be accepted if the null hypothesis is rejected. The null and alternative hypotheses are as follows:

> $\mathbf{H}_0$: All five rhetorical moves occur at least once in each discipline.
>
> $\mathbf{H}_\mathrm{A}$: All five rhetorical moves *do not* occur at least once in each discipline.

The hypotheses can be expressed algebraically for a five-move set as:

$$H_0 : \mathbb{D} | \forall m \{I, P, M, R, D\} \geq 1$$
$$H_a : \mathbb{D} | \forall m \{I, P, M, R, D\} \ngeq 1$$

where $\mathbb{D}$ is the set of ten disciplines, rhetorical move is represented by m and the set of rhetorical moves are I, P, M, R and D. For a more rigorous mathematical representation of the research problem, see A.1.

5.3.2 Presence of moves

Although the determination of moves could be considered subjective (as any classification task conducted by human evaluators could), the taxonomy and nomenclature of rhetorical moves used to classify the moves were based on an extensive review of the literature and pilot annotations of the corpus. A key consideration is the degree of granularity. Four-move and five-move systems are the most common taxonomies of rhetorical moves in research abstracts. The four-move taxonomy is INTRODUCTION MOVE, METHOD MOVE, RESULT MOVE and DISCUSSION MOVE (IMRD) while the five-move taxonomy is INTRODUCTION MOVE, PURPOSE MOVE, METHOD MOVE, RESULT MOVE and DISCUSSION MOVE (IPMRD). By extending the granularity to a five-move taxonomy the PURPOSE MOVE, which was previously a sub-move subsumed within the INTRODUCTION MOVE, is elevated to a full move. In addition, naming the RESULT MOVE is problematic, since in research studies that are neither empirical or experimental, an abstract artefact (e.g. an algorithm or proof) or concrete artefact (e.g. an autonomous robot assistant or self-driving car) is created.

The tagset used to annotate the corpus contained five rhetorical moves (INTRODUCTION MOVE, PURPOSE MOVE, METHOD MOVE, RESULT MOVE and DISCUSSION MOVE) and six sub-moves (BACKGROUND, GAP, PROBLEM, OVERVIEW, RESULT-AS-PRODUCT and METHOD-AS-PRODUCT). Although not the focus of this study, coding at one level of granularity deeper than reporting increases the veracity of the reported results.

TABLE 5.6: Presence-absence matrix for types of move[a] using five-move set by discipline[b]

Move	BOT	EC	IND	IP	IT	KDE	LING	MAT	MED	WC
Introduction	1	1	1	1	1	1	1	1	0	1
Purpose	1	1	1	1	1	1	1	1	1	1
Method	1	1	1	1	1	1	1	1	1	1
Results	1	1	1	1	1	1	1	1	1	1
Discussion	1	1	1	1	1	1	1	1	1	1

[a] 1 indicates presence, 0 indicates absence

[b] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing

A tailor-made script was created to extract the moves from the annotated corpus (See Appendix A.17). Regular expressions were used to discover the type of moves that appeared in each discipline. A presence-absence matrix was created to discover whether each move occurred in each discipline. Binary values (i.e. 1 for the presence and 0 for absence) were added to the matrix based on results of regular expression queries within each sub-corpus. As affirmed in presence-absence matrix for a five-move set of rhetorical moves shown in Table 5.6, the moves occurring in each discipline in the corpus were PURPOSE, METHOD, RESULT and DISCUSSION MOVES. The lack of an INTRODUCTION MOVE in the MED sub-corpus could be explained by the prescriptive structure that authors are required to adhere to. Medical abstracts tend to use structured abstracts (Hartley, 2003) which stipulate the subheadings to use

to make it easier for readers to search for and extract information from the abstract. This structure does not allow authors the flexibility to create additional sections. Although there is the possibility of including an INTRODUCTION MOVE in a different section, this did not occur in this corpus.

TABLE 5.7: Presence-absence matrix for types of move[a] using four-move set by discipline[b]

Move	BOT	EC	IND	IP	IT	KDE	LING	MAT	MED	WC
Introduction	1	1	1	1	1	1	1	1	1	1
Method	1	1	1	1	1	1	1	1	1	1
Results	1	1	1	1	1	1	1	1	1	1
Discussion	1	1	1	1	1	1	1	1	1	1

[a] 1 indicates presence, 0 indicates absence

[b] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing

The PURPOSE MOVE can be subsumed within the INTRODUCTION MOVE to create a four-move set of rhetorical moves, thereby adhering to the IMRD system, which many pedagogic books advocate. The presence-absence matrix for the four-move IMRD structure is shown in Table 5.6. The subsumption of the PURPOSE MOVE into the INTRODUCTION MOVE now means that the INTRODUCTION MOVE is now present in the MED sub-corpus. Thus, every move occurs in each discipline.

5.3.3 Presence of sub-moves

Sub-moves *per se* are not the focus of this sub-research question, but in order to gain greater insight into how the INTRODUCTION MOVES and RESULT MOVES are structured, sub-moves were also investigated.

TABLE 5.8: Presence-absence matrix for types of submove[a] by discipline[b]

Submove	BOT	EC	IND	IP	IT	KDE	LING	MAT	MED	WC
Background	1	1	1	1	1	1	1	1	0	1
Problem	1	1	1	1	1	1	1	1	0	1
Research gap	1	1	1	1	1	1	1	1	0	1
Overview	1	1	1	1	1	1	1	1	0	1
Method as result	0	1	1	1	1	1	0	0	0	1
Product as result	1	1	1	1	1	1	1	1	1	1

[a] 1 indicates presence, 0 indicates absence

[b] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing

As the INTRODUCTION MOVE is completely absent from the Medical discipline when using the five-move set, axiomatically none of the four sub-moves (BACKGROUND, PROBLEM, RESEARCH GAP and OVERVIEW) within the INTRODUCTION MOVE is present. Table 5.8 shows the presence-absence matrix for sub-moves. Merging the PURPOSE MOVE into the INTRODUCTION MOVE has no effect on the sub-moves and so

the five-move and four-move sets are not reported separately. It is noteworthy that information theory and the engineering discipline which includes industrial electronics and four other disciplines (EC, IP, KDE and WC) had the sub-move METHOD AS RESULT. The only sub-move occurring in each discipline was PRODUCT AS RESULT. All six sub-moves were present in six of the ten disciplines. These disciplines focus on the creation of new methods, five of which could be broadly grouped into information science (EC, KDE, IP, IT and WC). Therefore, research RESULT MOVES in these disciplines tended to be realised by the METHOD AS RESULT SUB-MOVE. The methods, however, may be tested and so an additional RESULT MOVE might be the PRODUCT AS RESULT SUB-MOVE, which could, for instance, be the results of a simulation test for a new method. The following examples are extracted from an abstract in the Information Theory sub-corpus. Example 9 shows the METHOD AS RESULT SUB-MOVE while Example 10 describes the results obtained by testing the method and so this move is classified as PRODUCT AS RESULT SUB-MOVE.

(9) In this paper, we present generic constructions of ZDB functions from functions with difference-balanced property.
METHOD AS RESULT SUB-MOVE, [IT 011]

(10) Employing these new ZDB functions, we obtain at the same time optimal (1) constant-composition codes, (2) constant-weight codes, and (3) perfect difference systems of sets, all with new and flexible parameters.
PRODUCT AS RESULT SUB-MOVE, [IT 011]

5.3.4 Sub-question 1 conclusion

As shown in 5.3.2, all five moves were present in nine disciplines while in the Medical discipline one move, namely the INTRODUCTION MOVE, was absent. Thus, for the five-move set, the null hypothesis $\mathbf{H}_0$ is rejected and the alternative hypothesis $\mathbf{H}_A$ is accepted. Medical abstracts tend to be structured (as they were in this corpus), which means that authors are required to follow a prescribed set of headings according to the type of research article. The guidelines for the British Medical Journal did not include a heading called Introduction, but included Purpose. There were instances of non-heading related moves occurring within different sections, but none included the INTRODUCTION MOVES within the section titled *Purpose*. It is likely that the shared background knowledge and mandated focus on purpose perhaps obviates the need for an INTRODUCTION MOVE.

However, when the PURPOSE MOVE is subsumed within the INTRODUCTION MOVE in the four-move set, each discipline contains each move, and so for the four-move set, the null hypothesis $\mathbf{H}_0$ is accepted.

Presence-absence matrices show whether a particular move occurred one or more times within a discipline. However, to gain a better insight into how each of these moves and sub-moves are used in each discipline, it is necessary to consider the frequency distribution of each of the rhetorical moves and sub-moves. The frequency

distributions using the five-move set of rhetorical moves are considered in Section 5.4.

5.4 Sub-question 2: Frequency of types of rhetorical moves

5.4.1 Preamble

This section addresses the second research question, namely:

> "How frequent is each move in research abstracts in each discipline?"

The null and alternative hypotheses are formulated as follows:

> $\mathbf{H}_0$: All five rhetorical moves occur in similar frequencies across all ten disciplines.
>
> $\mathbf{H}_A$: All five rhetorical moves *do not* occur in similar frequencies across all ten disciplines.

These hypotheses can be expressed algebraically as:

$$H_0 : \mathbb{D} | \forall m\{I, P, M, R, D\} = n \qquad \text{for } B \mid N$$
$$H_a : \mathbb{D} | \forall m\{I, P, M, R, D\} \neq n \qquad \text{for } B \mid N$$

where $\mathbb{D}$ is the set of ten disciplines, m represents move, and n represents the expected frequency. The expected frequency is based on one of two assumptions ($B \mid N$). Assumption B is where the frequency of each move in each discipline is balanced (i.e. constant) among the disciplines, i.e. approximately 20% for each of the five moves, or assumption N in which move frequency in each discipline is normally distributed.

To answer this research question, each of the distribution assumptions is considered in turn.

5.4.2 Balanced distribution

The raw frequencies of each move were counted for the whole corpus using a tailor-made script (See Appendix A.17 for script). Table 5.9 shows the raw frequencies for each move within the corpus and the percentage share by move.

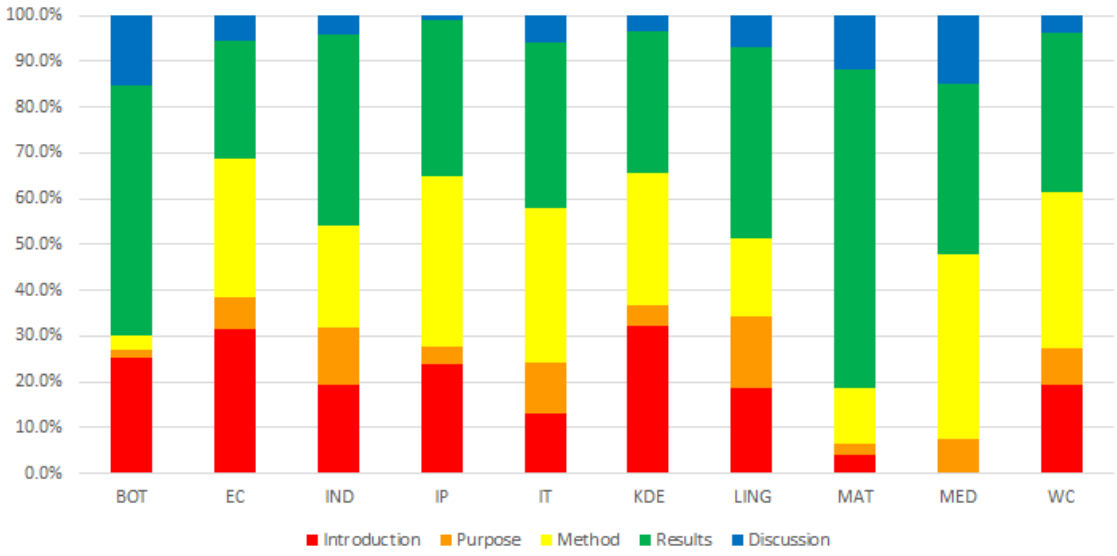
Both the PURPOSE MOVES and DISCUSSION MOVES are significantly less frequent than the INTRODUCTION, METHOD and RESULT MOVES. The arithmetic mean for the relative frequency of the PURPOSE MOVE for the corpus as a whole is 7.3% while the DISCUSSION MOVE stands at 7.8%. The mean for the INTRODUCTION MOVE is the third least frequent move at 18.5% while both METHOD MOVE (27.8%) and RESULT MOVE (38.6%) each comprise approximately one third of the total number of moves by percentage.

TABLE 5.9: Raw frequency distribution by type of move

Move	Raw frequency[a]	Percentage[b]
Introduction	1329	18.5%
Purpose	527	7.3%
Method	2005	27.8%
Results	2779	38.6%
Discussion	560	7.8%
Total	7200	100%

[a] Raw frequency is the exact number of instances occurring in the corpus
[b] Reported to one decimal place

Figure 5.4 provides a visual overview of the proportional distribution of rhetorical moves within each of the disciplinary sub-corpora. In this figure, the rhetorical moves are arranged in linear order from the INTRODUCTION MOVE through to the DISCUSSION MOVE, but this does not imply that that rhetorical moves occur in that order within abstracts. By noticing the difference in relative lengths of the coloured divisions representing different moves, it is clear that there is considerable variation among the frequencies of moves. The RESULT MOVE, coloured green accounts for over half of all the rhetorical moves in the BOT and MAT disciplines.



BOT = Botany; EC = Evolutionary computing; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; WC = Wireless computing

FIGURE 5.4: Frequency distribution of move types by discipline

Table 5.10 shows the raw frequencies of each move within each discipline in the corpus.The INTRODUCTION, METHOD and RESULT MOVES dominate the frequency count.

Table 5.11 shows the relative frequency by percentage for each move within each discipline in the corpus. The relative frequency enables easier comparison between and among disciplines as the effect of the considerable divergence that was caused by the difference in the length of abstracts among the disciplines is removed. The

TABLE 5.10: Raw frequency distribution for types of move[a] by discipline[b]

Move	BOT	EC	IND	IP	IT	KDE	LING	MAT	MED	WC
Introduction	208	240	127	175	81	257	104	11	0	126
Purpose	14	54	81	27	68	38	85	6	101	53
Method	28	230	146	274	211	231	94	32	536	223
Results	449	196	274	250	224	248	231	182	495	230
Discussion	127	42	27	8	36	27	39	31	199	24

[a] Raw frequency is the exact number of instances occurring in the corpus
[b] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing

remaining differences may, therefore, be considered a truer reflection of the genre of research abstracts for each discipline.

TABLE 5.11: Frequency distribution by percentage[a] for types of move[b] by discipline[c]

Move	BOT	EC	IND	IP	IT	KDE	LING	MAT	MED	WC
Introduction	25.2%	31.5%	19.4%	23.8%	13.1%	32.1%	18.8%	4.2%	0.0%	19.2%
Purpose	1.7%	7.1%	12.4%	3.7%	11.0%	4.7%	15.4%	2.3%	7.6%	8.1%
Method	3.4%	30.2%	22.3%	37.3%	34.0%	28.8%	17.0%	12.2%	40.3%	34.0%
Results	54.4%	25.7%	41.8%	34.1%	36.1%	31.0%	41.8%	69.5%	37.2%	35.1%
Discussion	15.4%	5.5%	4.1%	1.1%	5.8%	3.4%	7.1%	11.8%	15.0%	3.7%

[a] Reported to one decimal place
[b] Raw frequency is the exact number of instances occurring in the corpus
[c] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing

By analyzing Table 5.11 and Figure 5.4, different patterns of usage can be identified among the disciplines. Botany is a notable outlier with a low percentage for both PURPOSE MOVES (1.7%) and METHOD MOVES (3.4%). The lack of inclusion of the PURPOSE MOVE may be explained by the higher than average frequency of the INTRODUCTION MOVE. Unlike other disciplines, botany tends to embed the METHOD MOVE into the RESULT MOVE. This embedding may occur as subordinate clauses or through nominalization of the process, which Halliday (1985, p.358) terms grammatical metaphor. Example 11 and Example 12, both taken from the botany corpus, show RESULT MOVES with the METHOD MOVE embedded.

(11) Using a noninvasive microelectrode system, we showed that the BL sensor phototropin1 (phot1), the signal transducer NONPHOTOTROPIC HYPOCOTYL3 (NPH3), and the auxin efflux transporter PIN2 were essential for BL-induced auxin flux in the root apex transition zone.
RESULT MOVE [BOT 033]

(12) Using a novel differential RNA sequencing approach, which discriminates between primary and processed transcripts, we obtained a genome-wide map

> of transcription start sites in plastids of mature first leaves.
> RESULT MOVE [BOT 008]

In Example 11, the embedded METHOD MOVE contains far fewer word tokens (n = 5) than the RESULT MOVE (n = 32). The embedding of the METHOD MOVE tended to occur at either the start or end of the RESULT MOVE. As many of the procedures in botany follow standard operating practices, it is less necessary to explain the procedures in detail.

This is most likely because of the relative homogeneity of the collective methodological knowledge of the researchers. Unless a novel procedure is adopted, it is sufficient to name the methods and materials in procedure without any lengthy explanations.

The most striking feature for botany is that over half of the total number of moves are classified as RESULT MOVES (54.4%). Only materials science has a higher relative mean for RESULT MOVES at 69.5%. Both botany and materials science are natural sciences, and it appears that natural scientists value the dissemination of results, since research abstracts tend to adhere to the expected norms of the discourse community. It should also be noted that as the botany abstracts were graphical, the accompanying visual also conveyed information. Although not part of the formal annotation process, most of the accompanying graphics also showed or alluded to the results of the research.

The relative frequency of DISCUSSION MOVES for both botany and medicine is double the corpus mean of 7.3%. This may be related to the expectations of researchers in the life sciences. The impact of the research on the wider research community is stated explicitly. A corollary of this may be that the impact may not be apparent to researchers outside of a narrow research field within the respective disciplines. In contrast, research abstracts in the discipline of image processing rarely contain the DISCUSSION MOVE (n = 8, 0.9%). With just eight instances out of 844 moves, it appears that there is no need to show the wider implications of the research, presumably because discipline specialists are expected to be aware of them.

The range is the numerical difference between the lowest and highest values. For moves in general the range is the difference between the minimum percentage, which is 0.0% for INTRODUCTION MOVES in medical abstracts to 69.5% which is percentage of RESULT MOVES in materials science abstracts, and so the range is 69.5%.

Table 5.12 shows the minimum and maximum values by percentage for move frequency within disciplines. Range, or the difference between these values, is also provided. Within each discipline the range varies from 26.0% in Evolutionary Computing to 67.2% in materials science. The mean of the disciplinary ranges for the corpus is 37.8%.

The variance within the corpus can be measured using standard deviation. The standard deviation (SD) score shows the degree of variation from the mean. As there are only five variables, the expected mean for moves that are equally distributed is 20%). If move frequency follows a Gaussian (normal) distribution, SD calculates the

TABLE 5.12: Minimum, maximum and range by percentage[a] for frequency of move[b] types by discipline[c]

Move	BOT	EC	IND	IP	IT	KDE	LING	MAT	MED	WC
Maximum	54.4%	31.5%	41.8%	37.3%	36.1%	32.1%	41.8%	69.5%	40.3%	35.1%
Minimum	1.7%	5.5%	4.1%	1.1%	5.8%	3.4%	7.1%	2.3%	7.6%	3.7%
Range	52.7%	26.0%	37.7%	36.2%	30.3%	28.7%	34.7%	67.2%	32.7%	31.4%

[a] Reported to one decimal place
[b] Raw frequency is the exact number of instances occurring in the corpus
[c] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing

deviation from the mean which is a data point on the normal curve. The empirical rule specifies the amount of data expected to fall within each standard deviation away from the mean. Specifically, 68% falls within one standard deviation, 95% within two standard deviations and 99.7% within three standard deviations. Thus, with the assumption of a normal distribution divergence can be obtained by inspecting the shape of the curve.

The frequencies of move types are not balanced across the corpus. Thus, for Assumption *B* when frequency is expected to be constant, the null hypothesis is rejected, and the alternative hypothesis accepted.

5.4.3 Normal distribution

As the frequency values for each move vary, the frequency distribution is not balanced. Here the distribution is examined to see whether the assumption of a normal distribution holds. Figure 5.5 shows the plot of a curve with a standard normal distribution for reference.

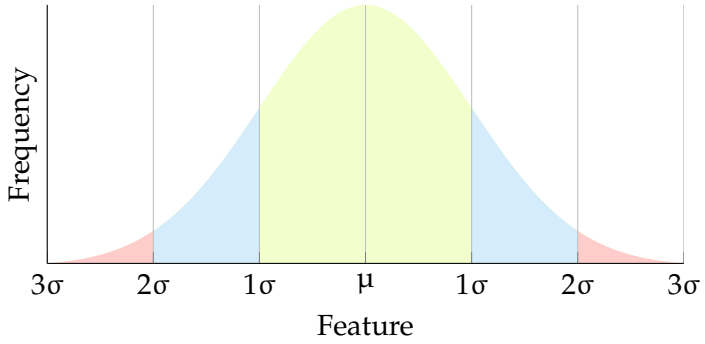


FIGURE 5.5: Standard normal distribution

Figure 5.6 shows the distribution of moves as a bar chart. It is apparent that the frequency of rhetorical moves across the corpus as a whole differ greatly. As the categorical variables do not have a fixed prescribed sequence, they could be positioned differently. If the categorical variables of the INTRODUCTION MOVE and PURPOSE MOVE are reversed as shown in Figure 5.7, the result appears to be a normal distribution curve albeit one that displays a degree of negative skew.

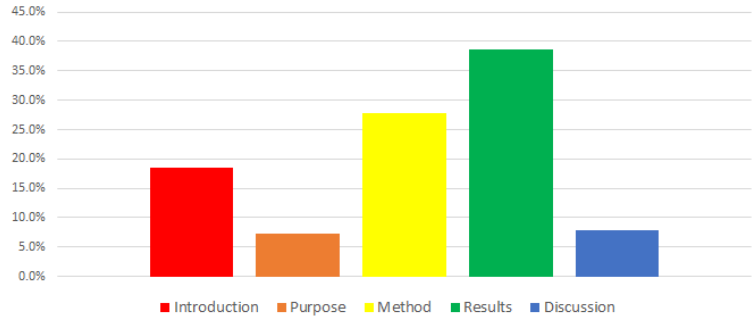


FIGURE 5.6: Frequency distribution in linear order (IPMRD)

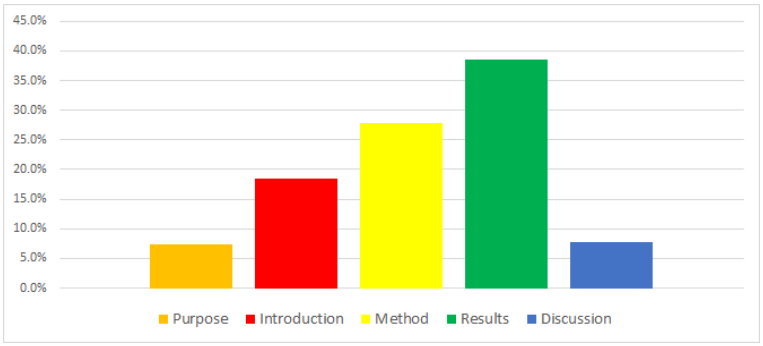
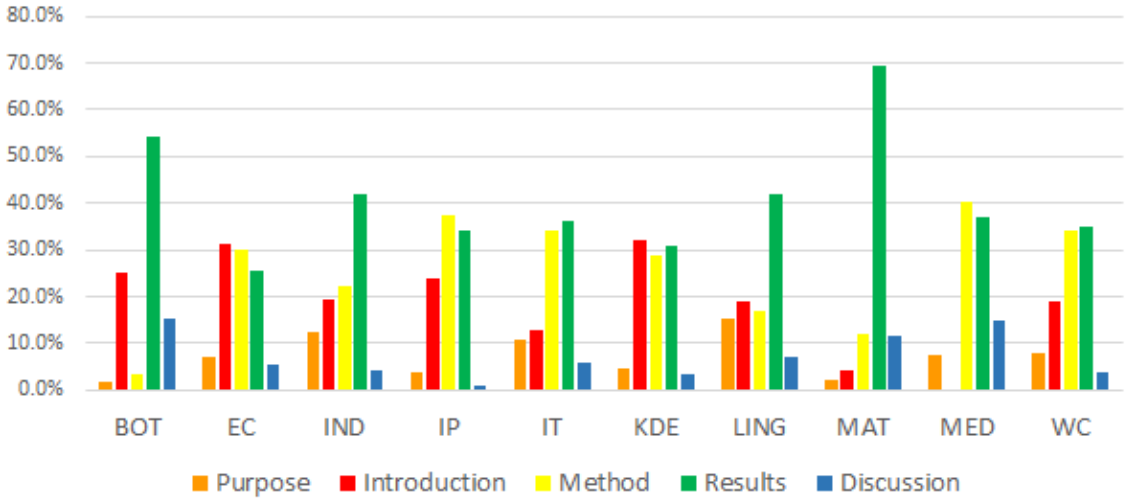


FIGURE 5.7: Reorganised frequency distribution starting with PURPOSE MOVE (PIMRD)

Thus, if the distribution of move types is relatively homogeneous across all ten disciplines, a similar normal-like distribution curve with a negative skew is expected when the sequence PURPOSE, INTRODUCTION, METHOD, RESULTS, DISCUSSION is used for the x axis.



[a] BOT = Botany; EC = Evolutionary computing; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; WC = Wireless computing

FIGURE 5.8: Bar chart of frequency distribution of move types by discipline starting with INTRODUCTION MOVE

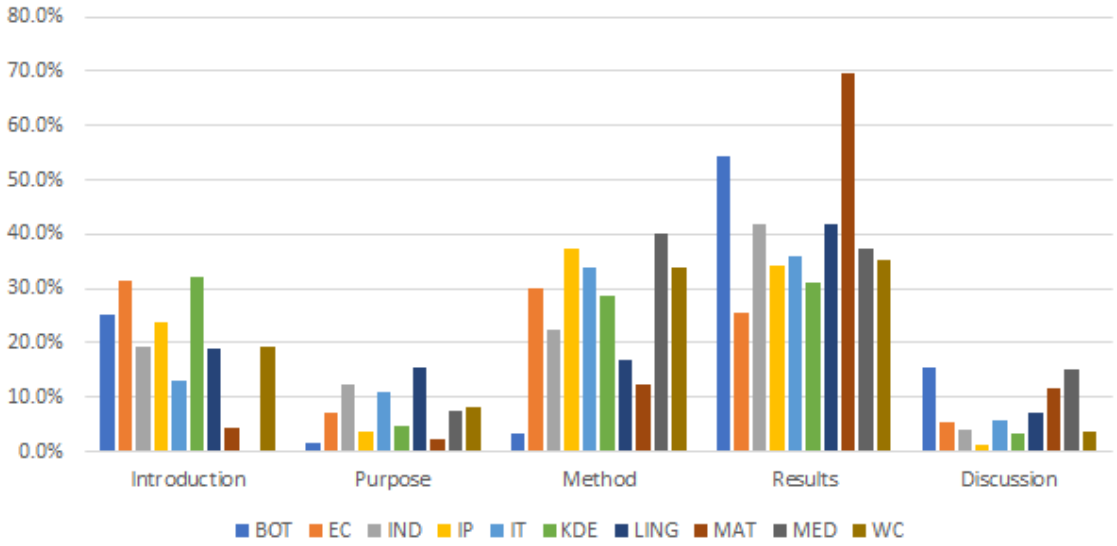
If the move types occur in a similar frequencies within each discipline then when the percentage frequency of each move is presented as a bar chart for each discipline, the same negative-skewed normal-like distribution curve would be expected. Figure 5.8 enables visual inspection of whether the proportional frequency distribution among disciplines is similar. This bar chart plots the moves in the sequence PIMRD rather than the expected linear order. As the bar charts for some disciplines do not follow the expected negatively-skewed normal-like distribution, it is clear that the relative frequencies of move types differ among disciplines. The research abstracts in the disciplines of Botany and Medical do not follow a normal curve using this order of variables, but should they be rearranged differently a normal-like curve can be obtained. To get a more precise indication, the frequencies of each move by discipline were listed in rank order starting from the least frequent and finishing with the most frequent. The most frequent move provides a strong indication of what is valued in the discourse community of the respective subject disciplines. Table 5.13 shows the frequency of moves in descending order for each discipline. The disciplines can be categorized into three groups based on the most frequent move. EC and KDE may be considered as Introduction-focused disciplines which need to situate readers in the research paradigm comparatively more frequently than other disciplines. IP, IT and MED may be considered more method-focused while the remaining disciplines could be considered more result-focused based on the proportion of move types.

TABLE 5.13: Rank frequency of move types[a] with each discipline[b]

Discipline[b]	1st	2nd	3rd	4th	5th
EC	I	M	R	P	D
KDE	I	R	M	P	D
IP	M	R	I	P	D
IT	M	R	I	P	D
MED	M	R	D	P	I
WC	R	M	I	P	D
IND	R	M	I	P	D
MAT	R	M	D	I	P
LING	R	I	M	P	D
BOT	R	I	D	M	P

[a] I = Introduction; P = Purpose; M = Method; R = Result; D = Discussion
[b] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing
[c] Moves are listed in descending order, i.e. 1st is the most frequent

Figure 5.9 enables closer comparison of the relative frequencies of individual move types among the ten disciplines. There appears to be a general trend that METHOD MOVE and RESULT MOVE make up the bulk of the moves while the PURPOSE MOVE and DISCUSSION MOVE are much less utilized.



[a] BOT = Botany; EC = Evolutionary computing; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; WC = Wireless computing

FIGURE 5.9: Bar chart of frequency distribution by move

5.4.4 Sub-question 2 conclusion

The frequency of moves is normally distributed across the whole corpus and within each discipline in the corpus although not for the same sequences of categorical variables; and, therefore, the frequencies are dissimilar. Thus, the null hypothesis $\mathbf{H}_0$ is rejected and the alternative hypothesis $\mathbf{H}_A$ is accepted.

5.5 Sub-question 3: Sequence of rhetorical moves

5.5.1 Preamble

This section presents the answer to the third sub-research question namely:

> "In what sequence do the moves occur in research abstracts in each discipline?"

This research question is exploratory in nature and so a hypothesis-testing approach is not undertaken. To answer this question, the specific sequences of rhetorical moves, i.e. move permutations, are identified, extracted and tabulated. Sequences can be classified into combinations and permutations. Combinations focus on the number of elements within a sequence but not on their order. Permutations, however, specify both the elements and their order. In this study, the permutations of rhetorical moves are under investigation. Sequences can be considered at multiple levels. In this study the following levels are considered:

1. the sequence in which two moves occur, e.g. the occurrence of adjacent pairs of moves

2. the sequence for all moves in a text, e.g. the permutations for research abstracts
3. the sequences that are realized out of a set of potential sequences, e.g. the actualized permutations versus potential permutations.

Each of these levels are considered in turn.

5.5.2 Occurrence of adjacent pairs of rhetorical moves

A combination of two moves MOVE A and MOVE B does not imply an order. However for any two sequential moves, there are two possible orders. Either MOVE A follows MOVE B, or MOVE B follows MOVE A. In either of these permutations MOVE A and MOVE B combine to form an adjacent pair of rhetorical moves. Table 5.14 shows the complete set of twenty adjacent pairs of moves that are possible using the five-move taxonomy of rhetorical moves, i.e. INTRODUCTION, PURPOSE, METHOD, RESULT and DISCUSSION. When the same rhetorical move occurs in sequential sentences, this is not counted as an adjacent pair as the sentences are merged into the same rhetorical move. Thus, two different moves are needed to create an adjacent pair of rhetorical moves.

TABLE 5.14: Set of adjacent pairs of rhetorical moves[a]

	Second move				
First move	**Introduction**	**Purpose**	**Method**	**Results**	**Discussion**
Introduction	–	IP	IM	IR	ID
Purpose	PI	–	PM	PR	PD
Method	MI	MP	–	MR	MD
Results	RI	RP	RM	–	RD
Discussion	DI	DP	DM	DR	–

[a] I = Introduction; P = Purpose; M = Method; R = Result; D = Discussion

With a four-move taxonomy, the number of adjacent pairs of moves falls to twelve permutations while the number increases to 30 adjacent pairs for a six-move taxonomy. Within the twenty potential permutations in the five-move taxonomy, some permutations would not expected to occur. For example, based on prior exposure to research abstracts, it seems unlikely that a DISCUSSION MOVE could occur before an INTRODUCTION MOVE. But, unlikely does not mean impossible, and so a tailor-made script was created to extract all the two-move permutations found in the annotated corpus (See Appendix A.14 for the script). Table 5.15 shows that adjacent pairs of move permutations were discovered for each permutation of moves.

When the same rhetorical move occurs in sequence, the moves are merged and so there are five cells for which results cannot be obtained. The results for the remaining cells were populated. Each potential permutation of adjacent pairs of rhetorical moves occurred at least once in the corpus. This diversity belies the simplicity of the IMRD model. For abstracts that strictly adhere to this move, only three adjacent pairs of moves, namely IM, MR and RD, are created. For abstracts that strictly adhere to the

TABLE 5.15: Presence-absence matrix for adjacent pairs of rhetorical moves

	Second move				
First move	**Introduction**	**Purpose**	**Method**	**Results**	**Discussion**
Introduction	–	1	1	1	1
Purpose	1	–	1	1	1
Method	1	1	–	1	1
Results	1	1	1	–	1
Discussion	1	1	1	1	–

extended five-move taxonomy of IPMRD, four adjacency pairs are created, namely IP, PM, MR and RD.

Yet, the actual corpus contained all 20 adjacent pairs of rhetorical moves. This number of adjacent pairs is 500% higher than would be expected for the five-move taxonomy. This result was rather unexpected. Prior to the corpus investigation, it was difficult to imagine how to write an abstract in such a way that, for example, the DISCUSSION MOVE could precede the METHOD MOVE or RESULT MOVE. The wide variety of permutations of adjacent pairs of moves reveals the complexity of rhetorical organization of research abstracts. This is in contrast to the over-simplistic models presented in the pedagogic literature. Novice writers of scientific articles may benefit initially from a prescriptive model to create an initial schema. However, given the wide variation found in the corpus, principles for the selection and sequencing of the rhetorical moves may be more useful than prescribing one sequence.

As the corpus of research abstracts comprises various disciplines, another search was conducted to investigate the occurrence of adjacent pairs of rhetorical moves within each discipline. This investigation aims to discover the degree of dispersion of adjacent pairs of moves throughout the corpus. A finer-grained presence-absence matrix was created for all disciplines. This matrix is shown in Table 5.16.

As shown in Table 5.16, the occurrences of adjacent pairs of moves among the disciplines outnumbers the absence of adjacent pairs. Evolutionary computing has the highest number of permutations at 18 actualized pairs of moves out of 20 potential pairs. Information theory, wireless communications and industrial electronics also displayed high numbers of adjacent pairs with 17 out of the possible 20. With a total of six pairs, medical research abstracts had the lowest number of different permutations of adjacent pairs. The number of adjacent pairs of moves was also comparatively low for materials science abstracts with a total of nine pairs occurring. The rhetorical organisation of medical abstracts most closely followed a linear IMRD structure with the caveat that the INTRODUCTION MOVE was replaced by the PURPOSE MOVE, creating the permutation PMRD.

TABLE 5.16: Presence-absence matrix for adjacent pairs of rhetorical moves[a] by discipline[b]

Adjacent pair	BOT	EC	IND	IP	IT	KDE	LING	MAT	MED	WC
IP	1	1	1	1	1	1	1	0	0	1
IM	1	1	1	1	1	1	1	1	0	1
IR	1	1	1	1	1	1	1	1	0	1
ID	1	1	0	0	0	1	1	0	0	0
PI	0	1	1	1	1	1	1	0	0	1
PM	1	1	1	1	1	1	1	1	1	1
PR	1	1	1	1	1	1	1	1	1	1
PD	0	1	1	0	1	0	1	0	0	1
MI	0	1	1	1	1	1	0	0	0	1
MP	0	1	0	1	1	0	1	0	0	1
MR	1	1	1	1	1	1	1	1	1	1
MD	0	1	1	1	1	1	1	1	0	1
RI	1	1	1	1	1	1	1	0	0	1
RP	0	1	1	1	1	1	0	1	1	1
RM	1	1	1	1	1	1	1	1	1	1
RD	1	1	1	1	1	1	1	1	1	1
DI	1	0	1	0	0	0	0	0	0	0
DP	0	1	0	0	0	0	0	0	0	0
DM	1	0	1	0	1	1	0	0	0	1
DR	1	1	1	1	1	1	0	0	0	1

[a] I = Introduction; P = Purpose; M = Method; R = Result; D = Discussion

[b] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing

5.5.3 Comparison of potential vs. actualized permutations of rhetorical move sequences for five-move scenarios

For two theoretical rhetorical moves (MOVE A and MOVE B), there are only two possible sequences, i.e. MOVE A then MOVE B or MOVE B then MOVE A (AB | BA). However, with the addition of each additional move the number of potential permutations increases rapidly. Permutations of moves show not only which rhetorical moves occur within an abstract, but the order in which the moves occur. Labov (1972) advocated measuring the frequency of variable occurrences against the frequency of opportunities for the variable to occur. This approach can be realized by calculating the number of theoretical permutations of rhetorical move sequences and comparing the potential permutations with the actualized permutations. The usual formula to calculate permutations is shown in Equation 5.1

$$ {}^{n}P_{k} = \frac{n!}{(n-k)!} \tag{5.1} $$

where n is the total number of the elements and k is the number of elements chosen.

Given the vast number of potential permutations, it is not possible to consider the complete set of potential permutation. Thus, three scenarios are considered for three-, four- and five-move permutations. In each permutation, the set of moves is specified, and each move is used once only with no repetitions. The datasets for the tables of potential permutations were automatically created using standard code that

harnesses Heap's algorithm (Heap, 1963) to generate the number of non-identical permutations when each move in each set is used once.

Tables 5.17, 5.18 and 5.19 show exact permutations possible in each scenario. With the addition of each successive rhetorical move, the number of permutations increases exponentially.

TABLE 5.17: Non-identical permutations for a three-move[a] (IMR) scenario

R-final	M-final	I-final
IMR	RIM	MRI
MIR	IRM	RMI

[a] I = Introduction; M = Method; R = Result

TABLE 5.18: Non-identical permutations for a four-move[a] (IMRD) scenario

D-final	R-final	M-final	I-final
IMRD	MIDR	DIRM	RMDI
MIRD	IMDR	IDRM	MRDI
RIMD	DMIR	RDIM	DRMI
IRMD	MDIR	DRIM	RDMI
MRID	IDMR	IRDM	MDRI
RMID	DIMR	RIDM	DMRI

[a] I = Introduction; M = Method; R = Result; D = Discussion

TABLE 5.19: Non-identical permutations for a five-move[a] (IPMRD) scenario

D-final	D-final	R-final	R-final	M-final	M-final	P-final	P-final	I-final	I-final
IPMRD	RIMPD	DIPMR	MDPIR	RDIPM	PRIDM	MRDIP	IMDRP	PMRDI	DPRMI
PIMRD	IRMPD	IDPMR	DMPIR	DRIPM	RPIDM	RMDIP	MIDRP	MPRDI	PDRMI
MIPRD	MRIPD	PDIMR	PMDIR	IRDPM	IPRDM	DMRIP	DIMRP	RPMDI	RDPMI
IMPRD	RMIPD	DPIMR	MPDIR	RIDPM	PIRDM	MDRIP	IDMRP	PRMDI	DRPMI
PMIRD	IMRPD	IPDMR	DPMIR	DIRPM	RIPDM	RDMIP	MDIRP	MRPDI	PRDMI
MPIRD	MIRPD	PIDMR	PDMIR	IDRPM	IRPDM	DRMIP	DMIRP	RMPDI	RPDMI
MPRID	PIRMD	PIMDR	IDMPR	IDPRM	DRPIM	DRIMP	RMIDP	RMDPI	MPDRI
PMRID	IPRMD	IPMDR	DIMPR	DIPRM	RDPIM	RDIMP	MRIDP	MRDPI	PMDRI
RMPID	RPIMD	MPIDR	MIDPR	PIDRM	PDRIM	IDRMP	IRMDP	DRMPI	DMPRI
MRPID	PRIMD	PMIDR	IMDPR	IPDRM	DPRIM	DIRMP	RIMDP	RDMPI	MDPRI
PRMID	IRPMD	IMPDR	DMIPR	DPIRM	RPDIM	RIDMP	MIRDP	MDRPI	PDMRI
RPMID	RIPMD	MIPDR	MDIPR	PDIRM	PRDIM	IRDMP	IMRDP	DMRPI	DPMRI

[a] I = Introduction; P = Purpose; M = Method; R = Result; D = Discussion

These tables show the number of permutations of moves for hypothetical three-move (IMR), four-move (IMRD) and five-move scenarios (IPMRD) in which each move is used once. It should be noted that these are just a few of many possible scenarios. For example, when considering a simple three-move scenario, there are nine other possible scenarios (e.g. IPM, IMD and IRD) that use different combinations of rhetorical moves selected from the five-move taxonomy. The actual situation is far more complex. For example, assuming the shortest abstract contains two of the

five moves and the longest contains all five moves (IPMRD), this gives a set of 320 potential permutations if no repetition of moves is permitted. If repetition of a move is permitted, the number of permutations increases dramatically to 3900 even when the total number of moves is capped at five.

To gain more insight into the variation of permutations, the potential permutations in Tables 5.17, 5.18 and 5.19 where compared with the permutations actualized in the corpus. Tables 5.20, 5.21 and 5.22 show the permutations discovered (represented by their code, e.g. IMR) while cells with zeros show the absence of those permutations in the corpus.

TABLE 5.20: Non-identical permutations for a three-move[a] (IMR) scenario discovered in corpus

R-final	M-final	I-final
IMR	RIM	0
0	IRM	RMI

[a] I = Introduction; M = Method; R = Result

TABLE 5.21: Non-identical permutations for a four-move[a] (IMRD) scenario discovered in corpus

D-final	R-final	M-final	I-final
IMRD	0	0	0
0	0	0	0
0	0	0	0
IRMD	0	0	0
0	0	IRDM	0
0	0	0	0

[a] I = Introduction; M = Method; R = Result; D = Discussion

TABLE 5.22: Non-identical permutations for a five-move[a] (IPMRD) scenario discovered in corpus

D-final	D-final	R-final	R-final	M-final	M-final	P-final	P-final	I-final	I-final
IPMRD	0	0	0	0	0	0	0	0	0
PIMRD	0	0	0	0	0	0	0	0	0
0	0	0	0	0	IPRDM	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

[a] I = Introduction; P = Purpose; M = Method; R = Result; D = Discussion

Although these three scenarios provide only a glimpse of the vast number of potential permutations, it is apparent that the frequency of opportunity far outweighs the frequency of occurrence. Within the three-move (IMR) scenario, it is notable that

each move was used in each of the three potential positions (initial, mid and final). Four out of the six potential permutations were discovered in the corpus. For the four-move (IMRD) scenario, three out of the 24 potential permutations were actualized. In each instance the INTRODUCTION MOVE, RESULT MOVE and DISCUSSION MOVE occurred in the same linear sequence. The position of the INTRODUCTION MOVE varied occupying spots immediately before and after the RESULT MOVE and in one case following the DISCUSSION MOVE. For the five-move (IPMRD) scenario, only three permutations were realized out of the potential 120 permutations. However, the notable features of two of the permutations are that each sequence begins with one of two moves INTRODUCTION MOVE and PURPOSE MOVE followed by the other. This two-move combination is followed by the sequence RESULT MOVE and DISCUSSION MOVE. As with the four-move scenario, the position of the METHOD MOVE varied. In two instances the METHOD MOVE was in the expected linear order before the RESULT MOVE, but in one instance, the METHOD MOVE followed the DISCUSSION MOVE.

5.5.4 Number of different permutations of rhetorical move sequences

To identify the permutations within each research abstract, a tailor-made script was created that extracted the permutations from the master dataframe (See Appendix A.18 for the script). This script collected a total of 196 permutations of rhetorical move permutations. Given that the oft-quoted IMRD and IPMRD are only two of these permutations, the pedagogic literature fails to recognise the 194 out of 196 permutations.

The first rhetorical move in a research abstract is particularly important. This is because readers often decide to read on or stop reading based on this first sentence. If all abstracts followed IRMD, the first move would always be the introduction. However, this is not the case. The first move may be considered as a focus move or initial-focus. That is the first move is used to focus the reader on what is important in the research. This is particularly pertinent for research in narrowly defined areas with large numbers of researchers who have to keep abreast of the latest developments.

An overview of the relative frequency of different permutations is given in Table 5.23. This table is organized by the initial move in a permutation, and so introduction-focused permutations are those that commence with an introduction move.

The organization of an abstract could be viewed as having two possible macro-organizing systems. The first is adherence to an envisaged, idealized or actual model sequence, which is some cases may be prescribed by an organization or publication. The second is rhetorical, in which the organization is purposeful and aims to persuade the reader. To grab the attention of a reader, particularly a reviewer, beginning an abstract with a hook that shows the strength of the research would be one rhetorical strategy. A primary purpose of research abstracts is to convince readers of the novelty, significance, substance and/or rigour of the research, and so the rhetorical move that is best able to fulfil this purpose may occupy the initial position. For research areas that are less widely studied, the INTRODUCTION MOVE could be used to show

TABLE 5.23: Number of permutations of rhetorical moves[a] commencing with a specific move

Initial-focus[b]	Raw number	Percentage [c]
Introduction-focussed	95	48.5%
Purpose-focussed	58	29.6%
Method-focussed	8	4.1%
Result-focussed	35	17.8%
Discussion-focussed	0	0%
Total	196	100%

[a] I = Introduction; P = Purpose; M = Method; R = Result; D = Discussion
[b] The initial move occupies a position of initial-focus
[c] Calculated to one decimal place

how the research fills a research gap or addresses a practical or theoretical problem. Studies that have compelling results, such as improvements in accuracy, speed or quantity may announce the results in the RESULT MOVE at the outset. A more subtle and commonly-utilized strategy is to propose a new method in a PURPOSE MOVE and then later in the abstract announce specific details in a RESULT MOVE. Research whose results have far ranging consequences could begin with the DISCUSSION MOVE to emphasize how the results generalize to other disciplines.

To gain a deeper understanding of the move permutations, the permutations for each of the five initial-focus collections of rhetorical move sequences were extracted and tabulated. The resultant permutations fell into four categories based on the initial move, namely INTRODUCTION MOVE, PURPOSE MOVE AND METHOD MOVE and RESULT MOVE. There were no instances beginning with the DISCUSSION MOVE.

There are 95 permutations beginning with an INTRODUCTION MOVE. The full list of introduction-first permutations is given in Table 5.24. As can be seen from the table, permutations vary from a solitary INTRODUCTION MOVE to a ten-move permutation of IRPMPMRMPR. Introduction-focused permutations were the most common accounting for approximately half (95 out of 196) of the total number of permutations. The frequency of the INTRODUCTION MOVE using as the first move could be used as an argument for the use of IMRD, since in most cases the INTRODUCTION MOVE is used in the initial position. However, the INTRODUCTION MOVE is followed by the METHOD MOVE in just 16 out of 95 instances, but is followed by the RESULT MOVE in 39 out of 95 instances. The introduction-focused permutations end with the DISCUSSION MOVE in 30 out of 95 instances.

Fifty-eight permutations began with the PURPOSE MOVE (see 5.25 for the full list). This is the second most common category of permutations based on initial-focus move comprising approximately 30% of all the permutations. Although there were no single-move permutations in this group, there were two two-move permutations, namely PURPOSE-METHOD and PURPOSE-RESULT. A ten-move permutation, PIRPMPMPMR, was also discovered in the purpose-first set of permutations. The PURPOSE MOVE was commonly followed by the INTRODUCTION MOVE (n = 18), PURPOSE MOVE (n = 20) and RESULT MOVE (n = 18) while it was rarely followed by the

TABLE 5.24: Permutations of rhetorical moves[a] beginning with INTRODUCTION MOVE

I	IPIPMRMPMR	IPRIR	IRMD
ID	IPIR	IPRIRMR	IRMDM
IM	IPIRM	IPRM	IRMDR
IMD	IPIRMR	IPRMR	IRMIM
IMI	IPM	IPRMRD	IRMIMR
IMPD	IPMD	IPRMRMR	IRMIR
IMPMR	IPMIRD	IPRMRPMR	IRMIRMR
IMR	IPMIRIR	IPRPMR	IRMR
IMRD	IPMIRMR	IR	IRMRD
IMRDR	IPMR	IRD	IRMRDMRD
IMRDRD	IPMRD	IRDIR	IRMRDR
IMRI	IPMRDRD	IRDIRD	IRMRDRD
IMRIMRD	IPMRI	IRDR	IRMRI
IMRIRD	IPMRMR	IRDRD	IRMRIR
IMRM	IPMRMRD	IRDRDIRD	IRMRM
IMRMIR	IPMRMRMR	IRI	IRMRMD
IMRMR	IPMRPR	IRIMR	IRMRMR
IMRMRPD	IPR	IRIMRMR	IRMRMRI
IP	IPRD	IRIPMR	IRMRMRMRD
IPD	IPRDM	IRIR	IRP
IPI	IPRDR	IRIRD	IRPMPMRMPR
IPID	IPRDRD	IRIRMD	IRPMR
IPIMR	IPRDRMRD	IRIRMR	IRPMRPD
IPIPMI	IPRI	IRM	

[a] I = Introduction; P = Purpose; M = Method; R = Result; D = Discussion

TABLE 5.25: Permutations of rhetorical moves[a] beginning with PURPOSE MOVE

PDIPM	PIPMRM	PM	PMRD	PR	PRMR
PDR	PIPRMR	PMD	PMRDR	PRD	PRMRD
PID	PIR	PMI	PMRDRMR	PRDMR	PRMRID
PIM	PIRD	PMIMI	PMRI	PRDR	PRMRM
PIMIR	PIRDR	PMIMR	PMRM	PRI	PRMRMR
PIMR	PIRMPDMD	PMIMRD	PMRMR	PRID	PRMRPMR
PIMRD	PIRMR	PMIR	PMRMRD	PRIR	PRPR
PIMRMR	PIRMRMD	PMIRDR	PMRMRM	PRIRM	PRPRD
PIPIR	PIRPMPMPMR	PMPMR	PMRPR	PRM	
PIPMR	PIRPR	PMR	PMRPRD	PRMDR	

[a] I = Introduction; P = Purpose; M = Method; R = Result; D = Discussion

DISCUSSION MOVE.

TABLE 5.26: Permutations of rhetorical moves[a] beginning with METHOD MOVE

MIPRMR	MRMI
MR	MRMR
MRD	MRMRD
MRDRD	MRMRMRD

[a] I = Introduction; P = Purpose; M = Method; R = Result; D = Discussion

There were comparatively few method-first permutations. In total, only eight permutations were discovered. The shortest permutation being METHOD-RESULT while the longest was MRMRMRD at seven moves. Of these permutations, the presence of MR and MRD were unsurprising. Cycling through adjacent pairs of moves occurred in five of the eight permutations. The METHOD-RESULT pair of moves was repeated in sequence in MRMR, MRMRD and MRMRMRD. The METHOD-RESULT was repeated with interruption in MIPRMR. In this permutation the IP moves were inserted between the first METHOD MOVE and the first RESULT MOVE. Cycling through moves also occurred for RESULT-DISCUSSION in the sequence MRDRD.

TABLE 5.27: Permutations of rhetorical moves[a] beginning with RESULT MOVE

R	RIPR	RMIM	RMRM
RD	RIRD	RMIMR	RMRMR
RDMIR	RIRM	RMIRM	RMRMRD
RDPMR	RIRMDR	RMIRMR	RMRMRMD
RDRD	RIRMR	RMR	RPR
RIM	RIRMRD	RMRD	
RIMR	RM	RMRDR	
RIMRD	RMD	RMRDRD	
RIMRDMR	RMDR	RMRI	
RIMRMR	RMI	RMRIMD	

[a] I = Introduction; P = Purpose; M = Method; R = Result; D = Discussion

Result-first permutations were surprisingly common. Thirty-five permutations were identified with the shortest being one move while two permutations comprised seven moves (RIMRDMR and RMRMRMD). Specialist informants noted that research abstracts, particularly for short research articles (also known as short communications and letters), tended to commence with a RESULT MOVE. The informants explained that this was because short research articles usually address known problems. The need to introduce the problem is therefore obviated. Writers that chose to front the RESULT MOVE are likely to be aiming to highlight the importance or significance of their results while those that begin with a METHOD MOVE might be trying to focus readers on their methodological rigour.

No permutations beginning with a DISCUSSION MOVE were found in this corpus. At times, the DISCUSSION MOVE occurred before the INTRODUCTION MOVE. Examples from the introduction-focused collection of permutations include IRDIR, IRDIRD and IRDRDIRD. But, this fronting of the DISCUSSION MOVE only occurred when

abstracts described the development and evaluation of an algorithm or system. The DISCUSSION MOVE related to the development phase while the INTRODUCTION MOVE related to the evaluation phase.

It is clear that the descriptive analysis of a corpus of published scientific abstracts has shown that the IMRD model does not reflect actual practice. The vast number of permutations that were discovered show that even among single scientific disciplines, there is great variation. Botany and Medicine were however notable exceptions. In the botany sub-corpus, the emphasis on the RESULT MOVE combined with the comparative brevity of the research abstracts reduced the number of permutations . The medical sub-corpus contained only structured abstracts with prescribed headings (Purpose, Method, Results, Discussion). This appeared to limit authorial choice as almost all authors strictly followed.

5.5.5 Sub-question 3 conclusion

This section has described and listed sequences in which moves occur in research abstracts in each discipline. The sequences were considered at the level of two moves. In this analysis, all twenty potential adjacent pairs of rhetorical moves were found. On further investigation, the disciplinary variation in the frequency of adjacent pairs of moves were identified. The disciplines BOT, MAT and MED showed the least variation with between six and thirteen adjacent pairs. The MED sub-corpus has the least variation with only six adjacent pairs due to its linear structure. However, engineering and information science disciplines (e.g. EC, IND, IP, IT, KDE and WC) showed the most variation with over 15 different adjacent pairs per discipline.

The potential permutations were compared with the actualized permutations in three scenarios. In each scenario only a few of the potential permutations were realized. The proportion of realized permutations decreased as the number of rhetorical moves increased. The actualized permutations are, therefore, inversely proportional to the number of moves in the potential permutation. Although only three conditions were considered, the results are likely to be similar for other scenarios.

In total, 196 permutations of rhetorical moves were discovered in this corpus. The situation is far more complex than described in the pedagogic literature. Given that the move sequence length ranged from one to ten and repetition of moves occurs, millions of permutations are possible. Yet, despite that number of potential permutations, just under 200 were realized. This indicates that factors are at play that constrain the selection of the sequence of moves. For the MED sub-corpus, the rigid prescribed PMRD structure is one obvious factor. For the MAT sub-corpus, the length limits the potential permutations. Longer abstracts have more flexibility in the positioning of moves, since there are more potential positions. Axiomatically, the converse is true for shorter abstracts.

5.6 Sub-question 4: Frequency of sequences of rhetorical moves

5.6.1 Preamble

This subsection presents the answer to the fourth sub-research question namely:

> "How frequent is each sequence in research abstracts in each discipline?"

To answer this question, frequency distribution of features has to be considered. Zipf's law (Bochkarev, Lerner, and Shevlyakova, 2014) explains the frequency distributions shown in many types of data. This empirical law states that rank order is inversely proportional to frequency. When shown graphically, a curve of exponential decay is created. However, when the same values are plotted using a logarithmic scale, a straight line is created. Zipf's formula is given in Equation 5.2.

$$f\left(k; N, s\right) = \frac{\frac{1}{k^s}}{\sum_{n=1}^{N} \frac{1}{n^s}} \tag{5.2}$$

where N is the number of elements, k is their rank, s is the value of the exponent characterizing the distribution.

Zipf's law is often used to describe the frequency distribution of words in a corpus. However, it is not known whether the frequency distribution of move permutations follows this law. It is likely, however, that for the 320 potential permutations in the simple scenario in Subsection 5.5.3, the Pareto principle applies. In short, a relatively small number of permutations dominate with most permutations being unrealized. However, what is not known is which particular permutations are highly frequent. According to the pedagogic literature, it is expected that IMRD should dominate. Section 5.5 showed the plethora of permutations far outnumbered those based on the simple linear IMRD or IPMRD models, but it could be that the frequency of the IMRD or IPMRD permutations outnumber the 195 other permutations.

The following hypotheses are formulated to test frequency distribution.

> $\mathbf{H}_0$: Zipf's law *does not* govern the frequency distribution of permutations of sequences of rhetorical move within the corpus.
>
> $\mathbf{H}_\mathrm{A}$: Zipf's law governs the frequency distribution of permutations of sequences of rhetorical move within the corpus.

These hypotheses can be expressed algebraically as:

$$\mathbf{H}_0 : \mathbb{D} | \forall p \neq f$$
$$\mathbf{H}_\mathrm{A} : \mathbb{D} | \forall p = f$$

where $\mathbb{D}$ is the set of ten disciplines, p is permutations of rhetorical move sequences and f represents a Zipfian distribution.

The frequency of sequences of rhetorical moves can be measured by considering:

1. the number of rhetorical moves in a sequence
2. the occurrence of adjacent pairs of rhetorical moves
3. the number of non-identical permutations of rhetorical moves

In this section the frequency distribution is investigated in all three ways to gain a more thorough understanding of how moves are sequenced. The hypotheses are tested when considering number of non-identical permutations of rhetorical moves.

The number of moves in move sequences are first investigated in Subsection 5.6.2. Adjacent pairs of rhetorical moves are considered in Subsection 5.6.3 while the number of non-identical permutations of rhetorical moves is addressed in Subsection 5.6.4.

5.6.2 Frequency of rhetorical moves sequences by number of moves

The common accounts of the rhetorical organization of research abstracts contain four moves (i.e. IMRD) or five moves (i.e. IPMRD). Therefore, the permutations of rhetorical move sequences are counted by the number of moves occurring in each permutation. The permutations were extracted and sorted by length calculated using the number of discrete moves. The results can be seen in Figure 5.10, which shows the frequency of the rhetorical move sequences by their length measured in individual moves.

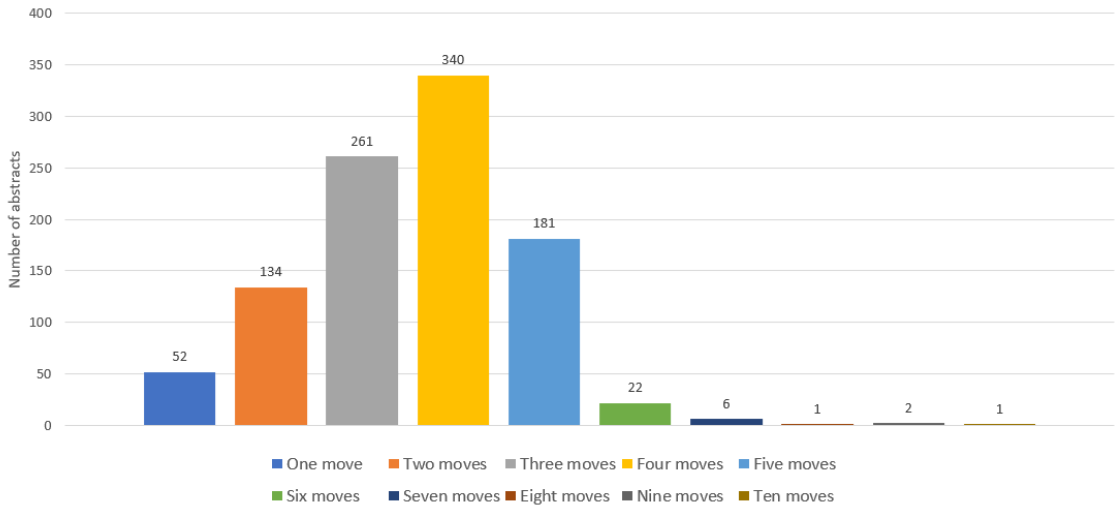


FIGURE 5.10: Number of moves in rhetorical move sequences

Four-move abstracts were most frequent (n = 340). Three-move abstracts, however, were the second-most common (n = 261) while five-move abstracts were the third-most frequent at 181 and two-move abstracts fourth-most frequent at 134. Single-move abstracts occurred in 152 instances. Abstracts with six moves or more were discovered, but there frequencies were far lower. Probably, the most surprising permutations were the single-move abstracts. These are not mentioned in any of the research literature on research abstracts, nor in the pedagogic literature targetting

writers of scientific research articles. Almost all comprised the RESULT MOVE, many of which were found in the materials science corpus.

5.6.3 Frequency of adjacent pairs of rhetorical moves

Any sequence comprises individual elements which are arranged in order. The order in research abstracts could be determined based on the previous or subsequent move. In Subsection 5.5.2, all twenty possible permutations of adjacent pairs of rhetorical moves were identified. The occurrence of an adjacent pair may be an oddity, and classed as an outlier, or it could be commonplace.

TABLE 5.28: Expected frequency of adjacent pairs of rhetorical moves[a]

	Second move				
First move	**Introduction**	**Purpose**	**Method**	**Results**	**Discussion**
Introduction	–	High	Very high	Low	Low
Purpose	Very high	–	High	Low	Low
Method	Low	Low	–	Very high	Low
Results	Very low	Very low	Low	–	Very high
Discussion	Very low	Very low	Very low	Very low	–

Table 5.28 shows the expected likelihood of the particular sequences of adjacent pairs of rhetorical moves. This table was created using the knowledge gained from the literature review and exposure to research abstracts in general, but without querying the corpus of scientific research abstracts.

To determine whether the expected usage matched the actual usage, a tailor-made script was created to count each adjacent pair of rhetorical moves (See Appendix A.5 for script). The number of occurrences of each adjacent pair was counted and tabulated (see Table 5.29).

TABLE 5.29: Actual frequency of adjacent pairs of rhetorical moves

	Second move				
First move	**Introduction**	**Purpose**	**Method**	**Results**	**Discussion**
Introduction	–	173	102	325	6
Purpose	50	–	297	144	7
Method	28	11	–	655	32
Results	53	18	352	–	357
Discussion	4	1	9	42	–

Table 5.29 shows the actual frequency of adjacent pairs within the corpus. The adjacent pairs of rhetorical moves with the highest frequency are:

1. METHOD-RESULT (n = 655)
2. RESULT-DISCUSSION (n = 357)
3. RESULT-METHOD (n = 352)
4. INTRODUCTION-RESULT (n = 325)

The first two adjacent pairs of moves fit IMRD model. However, the next two adjacent pairs do not match the model. Given that the RESULT-METHOD pair is almost as frequent as Result-Discussion, a more accurate model is needed. The fourth most common adjacent pair of INTRODUCTION-RESULT is particularly notable since the expected METHOD MOVE is omitted. To determine whether these results were specific to particular disciplines, a table of the frequency of adjacent pairs of rhetorical moves for all disciplines was tabulated.

TABLE 5.30: Adjacent pairs of rhetorical moves[a] by discipline[b]

Adjacent pair	BOT	EC	IND	IP	IT	KDE	LING	MAT	MED	WC
IP	14	37	26	14	14	22	31	0	0	15
IM	15	14	7	17	11	8	12	3	0	15
IR	74	37	19	46	25	70	12	7	0	35
ID	1	1	0	0	0	2	2	0	0	0
PI	0	8	9	4	8	8	7	0	0	6
PM	1	35	38	10	30	11	39	1	100	32
PR	13	8	29	11	20	12	31	5	1	14
PD	0	1	1	0	3	0	1	0	0	1
MI	0	3	1	5	5	7	0	0	0	7
MP	0	3	0	3	3	0	1	0	0	1
MR	26	77	66	79	68	78	59	24	101	77
MD	0	7	3	1	10	4	1	1	0	5
RI	5	6	10	15	4	8	1	0	0	4
RP	0	3	4	2	4	1	0	1	1	2
RM	7	50	31	71	51	75	8	3	1	55
RD	103	25	20	7	18	12	32	29	98	13
DI	3	0	1	0	0	0	0	0	0	0
DP	0	1	0	0	0	0	0	0	0	0
DM	1	0	2	0	2	3	0	0	0	1
DR	17	4	7	2	2	3	0	0	0	7

[a] I = Introduction; P = Purpose; M = Method; R = Result; D = Discussion

[b] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing

Table 5.30 shows the actual frequency of the particular sequences of adjacent pairs of rhetorical moves by each of the ten disciplines. However, with over 200 values, it is difficult to discern patterns within the table. To more easily visualize these results notice patterns, heat maps are provided in Figure 5.11. These heat maps colorize adjacent pairs of rhetorical moves by frequency for each discipline. The more frequent occurrences are coloured deeper and darker.

There are two patterns that stand out in the heat maps. First, the regularity of adjacent pairs in the medical corpus; and, second, the high frequency of MR and RD adjacent pairs. The regularity of medical abstracts is explained by their specified structure. Medical abstracts follow a prescribed structure, and almost every author (99 out of 100) adhered to the structured abstract format.

The adjacent pairs of MR and RD both include the RESULT MOVE. It appears that the RESULT MOVE is the most important move in scientific research abstracts. Discussions with specialist informants in information science (EC, IP, IT, KDE, WC) and engineering (IND) confirmed the primacy of the RESULT MOVE in those disciplines.

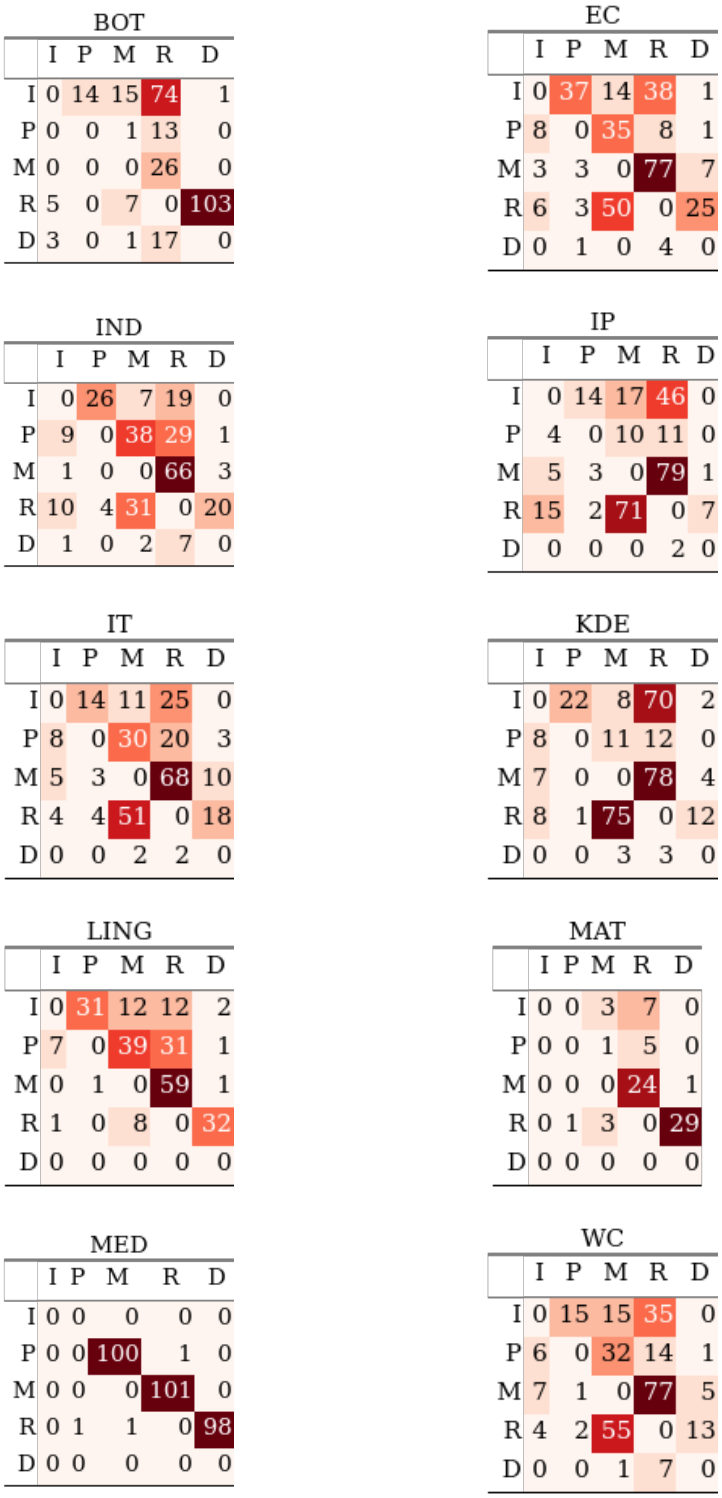


FIGURE 5.11: Heat maps for adjacent pairs of moves
[a] BOT = Botany; EC = Evolutionary computing; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; WC = Wireless computing

One journal editor explained that without a RESULT MOVE in the research abstract, the corresponding article would be unlikely to be accessed, downloaded or cited.

5.6.4 Frequency of permutations of rhetorical move sequences

Permutations of moves were extracted from the annotated corpus using a tailor-made script. Table 5.31 shows permutations of moves that had frequency counts of ten or more.

TABLE 5.31: The most frequent permutations of rhetorical moves[a] in corpus

Permutations	Number	Permutations	Number
PMRD	110	RD	25
IRMR	87	IPRD	22
PMR	61	PR	20
IR	52	MR	17
IRD	52	IMRD	14
R	50	IPMRD	12
RMR	36	IRMRD	11
IPMR	29	PMRMR	11
IPR	29	IRM	10
IMR	25	RM	10

[a] I = Introduction; P = Purpose; M = Method; R = Result; D = Discussion

These twenty permutations accounted for over two-thirds (n = 683) of all the move sequences. The millions of potential permutations were reduced to approximately 200 actualized permutations, which has in turn been reduced to twenty permutations that account for the lion's share of the permutations. Within these twenty permutations, just six permutations account for 40% of the permutations discovered.

The permutation PMRD was particularly common. This can be explained by the prescribed structure that medical abstracts were expected to adhere to. The permutation IRMR was the second most frequent at 87 instances. This permutation was used in research involving the development and evaluation of a method, system or algorithm. The first RESULT MOVE described the method, system or algorithm while the second RESULT MOVE produced numerical evidence showing its superiority compared to other methods, systems or algorithms. The permutation PMR was the third-most frequent. In this permutation, the PURPOSE MOVE introduced the overarching aim of the research while the RESULT MOVE provided specific details of what was achieved. The next two permutations (IR, IRD) tended to focus on results but provided sufficient background information to orientate readers to the discipline area. The single-move permutation, R, almost invariably occurred in the materials science corpus. Abstracts in the MAT sub-corpus are graphical and were considerably shorter than other disciplines. Given the limited space, researchers tended to emphasize results often ending up with abstracts comprising of a single RESULT MOVE.

Table 5.32 provides the list of the most frequent permutations of rhetorical moves for each discipline. The criteria for inclusion in the list was that the permutation occurred at least four times in the same disciplinary sub-corpus.

TABLE 5.32: Most frequent permutations[a] of rhetorical moves[b] by discipline[c]

BOT	No	**EC**	No	**IND**	No	**IP**	No	**IT**	No
IRD	35	IRMR	11	PMR	16	IRMR	23	PMR	8
IMRD	10	IPMR	9	IPR	6	RMR	15	IR	5
IR	10	IPMRD	6	PMRMR	4	IR	7	IRMR	5
IPRD	9	IR	6	PR	4	IMR	4	RMR	5
IRMRD	4	PMR	5	RMR	4			RMD	4
IRDRD	4	IMR	4						
		RMR	4						
KDE	No.	**LING**	No	**MAT**	No	**MED**	No	**WC**	No
IRMR	31	PMR	17	R	41	PMRD	97	IRMR	11
IR	8	IPR	11	RD	20			PMR	10
IPMR	4	PMRD	8	MR	15			IR	7
		PR	8	MRD	5			RMR	4
		IMR	7	IR	4				
		IPMR	7						
		IPRD	5						
		PRD	5						
		IPMRD	4						
		IRD	4						

[a] inclusion based on threshold frequency of 4
[b] I = Introduction; P = Purpose; M = Method; R = Result; D = Discussion
[c] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing

The MED sub-corpus, as has been mentioned earlier, is the most homogeneous, closely adhering to the prescribed permutation PMRD. The frequent move permutations in the MAT sub-corpus are the shortest ranging from one to three moves. As noted earlier, this is most likely due to the comparatively low number of words. According to the instructions for authors, the abstract should be less than or equal to 200 words and should state the key findings. Apart from emphasizing that the abstract is the most important selling point of the manuscript and the inclusion of the RESULT MOVE, there is no recommendation to include any other moves.

According to specialist informants, researchers in some disciplines, notably materials science and botany tend to use boilerplate templates to create their abstracts. In materials science and wireless communication, it might be that only the names of the materials, method and values are swapped out, but the overall structure is kept. In botany, researchers frequently stated that they simply included their most important result and added the INTRODUCTION MOVE, METHOD MOVE or DISCUSSION MOVE when space permitted or the information was central to their claim. The METHOD MOVE is only present in two out of six permutations for a total of 14 out of the 72 instances of frequent permutations. This shows the lack of necessity for a method move. The narrow field of research with highly established and well-known methods

may obviate the need to report the method within the very limited word length allocated for graphical abstracts.

The linguistics corpus is interesting because there are ten frequent permutations of moves, which is almost 50% more than any of the other disciplinary corpora. Abstracts in the BOT and EC disciplines had six and seven frequent permutations respectively. The variation in the rhetorical organization in linguistics may be related to the discipline itself. Few linguists would claim to rigidly follow boilerplate templates in the same was as MAT and BOT researchers, and most would be confident enough not to rely on a template.

The non-linear palindrome permutation of RMR occurs in the most frequent permutations for five disciplines (EC, IND, IP, IT, WC). All five disciplines regularly include research that involves the reporting of the creation of a method, algorithm or system and its subsequent evaluation.

5.6.5 Sub-question 4 conclusion

The frequency distribution was investigated three ways, namely by:

1. the number of moves per permutation,
2. the occurrence of two-move adjacent pairs of rhetorical moves, and
3. the number and frequency of different permutations of rhetorical moves.

The null hypothesis to be tested was that Zipf's law *does not* govern the frequency distribution of permutations of sequences of rhetorical move within the corpus.

When the number of moves per permutation were considered, the distribution of moves when aligned in rank order, dropped linearly and not logarithmically. Thus, the null hypothesis is accepted for this scenario.

When the frequency distribution of adjacent pairs of rhetorical moves were placed in rank order and graphed, the distribution of moves when aligned in rank order dropped following a sigmoid curve, or an s-curve. The distribution was not logarithmic and so Zipf's law was not obeyed. Thus, the null hypothesis is accepted for this scenario as well.

The final scenario considers the frequency of the permutations by abstract. The logarithmic graph for this distribution does not follow a straight line and so the distribution does not follow Zipf's law. Thus, the null hypothesis is accepted for this scenario, and so for each scenario, Zipf's law does not describe the frequency distribution of move permutations.

The Pareto principle appears to hold with the majority of the frequency distribution being governed by a minority of the permutations. The Pareto principle states that approximately 80% of the frequency distribution is governed by 20% of the features. When considering the number of moves per permutation, the most common two values accounted for 60% of the distribution. At the level of adjacent pairs of moves, the most frequent three adjacent pairs of moves accounted for slightly over

50% of all the adjacent pairs. Finally, at the level of move sequence permutations by abstract, 20 permutations out of almost 200 accounted for over two-thirds of all the move sequences.

5.7 Sub-question 5: Similarities in rhetorical organization

5.7.1 Preamble

This section presents the answer to the fifth sub-research question namely:

> "What are the similarities in rhetorical organization between the disciplines?"

This research question is hypothesis-generating rather than hypothesis-testing, and so an exploratory approach is adopted. While analyzing the corpus and the permutations of rhetorical moves, three categories of features stood out. These categories will be referred to as *dimensions*. The three dimensions are linearity, cyclicity and variation. Each of these dimensions is discussed in depth in Subsection 5.7.2. Using these three dimensions as a base, a framework of three interlocking circles (henceforth referred to as *Borromean rings*) was created as the vector space onto which each corpus of abstracts could be mapped. Subsection 5.7.3 presents the *Borromean Rings* framework using basic set theory, and shows the location of each sub-corpus within the framework.

5.7.2 Three distinctive dimensions

The three distinctive dimensions were identified over the course of this study. At the outset, none of these dimensions had been considered. The dimensions were most noticeable when using the *Move Highlighter* which colorizes rhetorical moves. This colorization made it easy to notice when abstracts deviated from the expected order. This realization meant lead to the consideration of the linearity dimension. In tandem with linearity, reoccurring colour sequences were found using the *Move Highlighter*. Pairs, or at times trios, of moves were repeated within an abstract. This discovery initially led to the concept of recursivity or recursion. But, in order, to avoid criticism from computer scientists who use recursivity with a more specific meaning, the easier-to-follow term cyclicity was selected. The final dimension aims to distinguish disciplines by the degree to which writers follow established permutations or create novel permutations. Unlike the previous two dimensions, this dimension applies to the corpus or dataset as a whole and not to any individual research abstract since it is impossible to ascertain a value for variation from a single instance.

Through the hands-on experience of working with the texts in the corpus, and analyzing the permutations manually and automatically, using queries and functions to identify patterns within the permutations, these three distinctive dimensions were confirmed to be pervasive. Each of the three dimensions of linearity, cyclicity and variation are discussed in turn.

Linearity dimension

Many genres have discrete beginnings, middles and ends. Informal and formal letters begin with greetings, before stating the message and finishing with a salutation. Academic essays tend to follow a tripartite structure of introduction, body and conclusion. Each of which may be further subdivided. Thus, the linearity of letters could be represented as greeting-message-salutation and academic essays could be represented linearly as introduction-body-conclusion. Research abstracts are prescriptively described in the pedagogic literature as IPMRD. By using this five-move order as the default for linear abstracts, abstracts in which a move occurs out of sequence are classified as non-linear. For example, given that the PURPOSE MOVE is expected before the METHOD MOVE, permutations with RESULT MOVE occurring before the PURPOSE MOVE are classified as non-linear. Teachers of academic and research writing with backgrounds in the humanities and little exposure to scientific writing may, based on their experience, expect scientific research abstracts to adhere to the linear IMRD structure despite disciplinary expectations to the contrary. All disciplines contained abstracts that were linear, but some disciplines contained abstracts that were non-linear.

Cyclicity dimension

Cyclicity is inversely correlated to linearity. Abstracts can show linearity or cyclicity, but cannot show both. Non-linear abstracts may contain cyclicity (e.g. IMRMRD, RMRMD), but they may be non-linear and non-cyclical (e.g. RM, RMI). Cyclicity has previously been revealed in move analyses of introductions using the Swalesian CARS model (Anthony, 1999; Bunton, 2014). Bunton (2014) in his study of doctoral dissertations showed that a modified CARS model was used cyclically. Swales (2004) also modified the CARS model by providing an option to cycle through move 1 (establishing a territory) and move 2 (establishing a niche). One of the effects of cycling through moves is that the subsequent adjacent pair of moves may be non-linear. For example, cycling through the INTRODUCTION MOVE and METHOD MOVE of IMRD creates the permutation IMIMRD. In this permutation, the second adjacent pair of rhetorical moves becomes MI. In fact, the METHOD MOVE is related to the first INTRODUCTION MOVE, and the second INTRODUCTION MOVE relates to second METHOD MOVE.

Variation dimension

Linearity and cyclicity can be discovered in a single abstract or by investigating a corpus. Variation, however, cannot be determined through a single abstract since one abstract can only provide one permutation, from which the degree of variation cannot be deduced or induced. The variation in permutations can only be measured by counting the number of non-identical permutations within the target set of texts. Disciplines that display more homogeneity most likely are written for an audience

of readers who share a similar knowledge base and the same set of disciplinary expectations. Hyland (2004, p.75) explains that variation occurs as writers try to frame their work in way that can "convince others of their work, given the particular circumstances of their research". The variation in the patterns of move sequences is an indication of choices that writers use to convince members of their own community of practice of the value of their research.

5.7.3 *Borromean Rings* framework

The *Borromean Rings* framework is the result of applying set theory to discover the potential categories that the different combinations of the three dimensions may create. These categories are mapped two-dimensionally onto a Venn diagram. A three-set Venn diagram that looked like *Borromean rings* was selected as the framework or vector space.

Set theory

To attempt to discover a pattern in these three dimensions in research abstracts, set theory was used to identify the possible combinations of three dimensions in order to decide how many regions are needed to create the base for a framework. Given that there are three binary variables, the total number of combinations is the product of the number of possible values for each variable:

$$2x2x2 = 8 \tag{5.3}$$

and so abstracts could theoretically be classified into one of the following eight combinations, which are:

1. Variation only (V)
2. Linearity only (L)
3. Cyclicity only (C)
4. Both Variation and Linearity ($V \cap L$)
5. Both Linearity and Cyclicity ($L \cap C$)
6. Both Cyclicity and Variation ($C \cap V$)
7. Variation, Linearity and Cyclicity ($V \cap L \cap C$)
8. No Variation, Linearity nor Cyclicity ($V^c \cap L^c \cap C^c$)

Therefore, there are three regions with one dimension, three regions with two dimensions, one region with all three dimensions and one region with no dimensions. These eight categories represent all the theoretical combinations. However, given that it is impossible for a discipline to show both linearity and cyclicity, this region will be an empty set.

Venn diagram

To create a two-dimensional representation of the three variables, a Venn diagram was created as shown in Figure 5.12. Venn diagrams are commonly used in set theory to map the relationship of sets, Venn diagrams are also a tool used in propositional logic to solve decisions or to evaluate the validity of arguments. In this case, the label of each circle could be considered to be a proposition stating that a particular dimension exists in the dataset or corpus.

Within the area of the three-interlocking circles or *Borromean rings*, there are seven regions while the area outside the circles is the eighth region. The characteristics of each disciplinary corpus of research abstracts can be plotted into this three-set Venn diagram based on the dimensions of variation, linearity and cyclicity. If each ring represents the proposition that there exists the specified dimension, then it is possible to map the dimensions of each corpus onto the Venn diagram. Given that the variation dimension is only recognizable at the level of corpus or dataset, it is not possible to use the *Borromean Rings* to map individual abstracts.

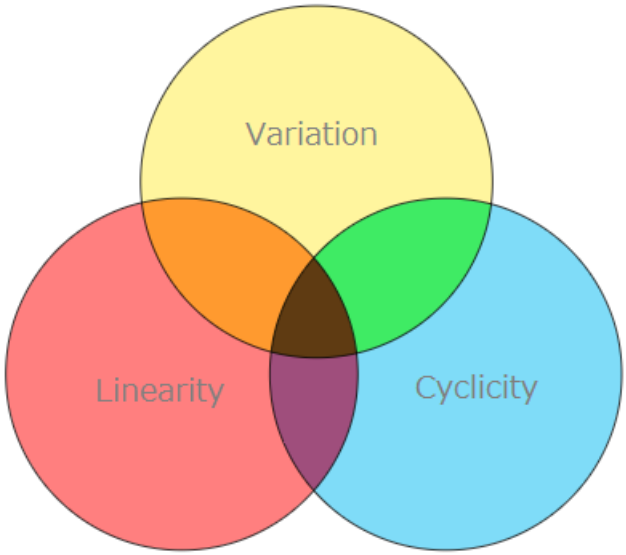


FIGURE 5.12: Venn diagram comprising *Borromean Rings*

Table 5.33 shows which elements can be mapped to which region. For example, consider a corpus of research abstracts which show both cyclicity and variation. This discipline is mapped to the green intersection whose elements are members of both the variation set and cyclicity set.

Each region within the *Borromean Rings* framework can be represented by a three-digit binary number. Figure 5.13 shows the *Borromean Rings* framework labelled with binary numbers. Each digit in the binary number indicates the presence or absence of one of the three dimensions: variation, linearity and cyclicity. When the digit is 1, the dimension is present. When the digit is 0 the dimension is absent. The first digit of the left represents linearity, the second cyclicity and the third variation. Thus,

TABLE 5.33: Notation and regions associated with combinations of the three features

Feature(s)	Notation[a] [b]	Colour of region[c]
Variation only	V	yellow
Linearity only	L	red
Cyclicity only	C	blue
Both variety & linearity	$V \cap L$	orange
Both linearity & cyclicity	$L \cap C$	purple
Cyclicity & variation	$C \cap V$	green
Variation, linearity & cyclicity	$V \cap L \cap C$	brown
Neither variation, linearity nor cyclicity	$V^c \cap L^c \cap C^c$	white

[a] The symbol $\cap$ indicates an intersection between two regions
[b] The superscript c as in X^c indicates the complement, i.e. elements that are not members of a set
[c] Colour of the regions in the Borromean rings

the code 111 represents the presence of all three dimensions and so it occupies the intersection of three rings coloured in brown while 100 represents the presence of linearity and the absence of both cyclicity and variation. This digit describes the region coloured red.

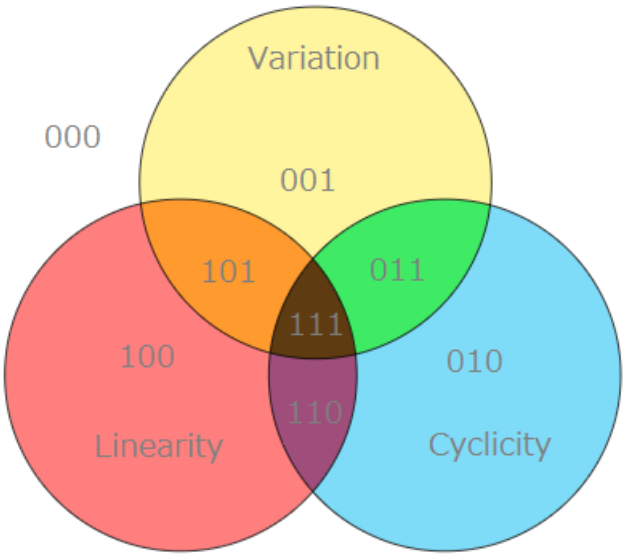


FIGURE 5.13: *Borromean Rings* framework labelled with binary codes

Truth tables

In order to map disciplines onto this framework, truth tables are used to evaluate the presence (and therefore truth) or absence (and therefore falsity) of each dimension in turn. Based on the results of the evaluation of the three dimensions, a three-digit binary code is created that can be used to map the disciplinary corpora of research abstracts onto the most appropriate region in the framework. Table 5.34 shows a truth

table that evaluates the overall presence or absence of a dimension throughout each disciplinary corpus.

TABLE 5.34: Truth table[a] for rhetorical organization of research abstracts by discipline[b]

Feature	BOT	EC	IND	IP	IT	KDE	LING	MAT	MED	WC
Linearity	1	0	0	0	0	0	1	1	1	0
Cyclicity	0	1	1	1	1	1	0	0	0	1
Variation	1	1	1	1	1	1	1	0	0	1
Binary code	101	011	011	011	011	011	101	100	100	011
Region colour	orange	green	green	green	green	green	orange	red	red	green

[a] 1 represents the presence of a feature while 0 represents the absence.
[b] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing

The resultant three-digit binary code is then plotted in the appropriate vector space. The results of the plot of the results extracted from Table 5.34 are shown in Figure 5.14.

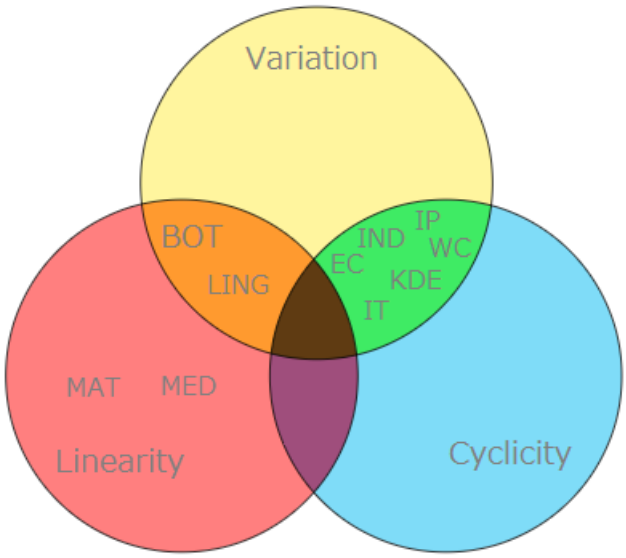


FIGURE 5.14: Disciplines mapped onto the *Borromean Rings* framework

5.7.4 Multidimensional scaling (MDS)

Section 5.7.3 proposed a *Borromean rings* framework onto which disciplines can be mapped based on the dimensions discovered in a corpus of research abstracts. To check the veracity of this framework, multidimensional scaling (MDS) is used to reduce the three dimensions of linearity, cyclicity and variation to two dimensions. After which a k-means clustering algorithm is used to group the disciplines using Euclidean distance (See Subsection 4.5.4 for a detailed description of MDS and k-means clustering).

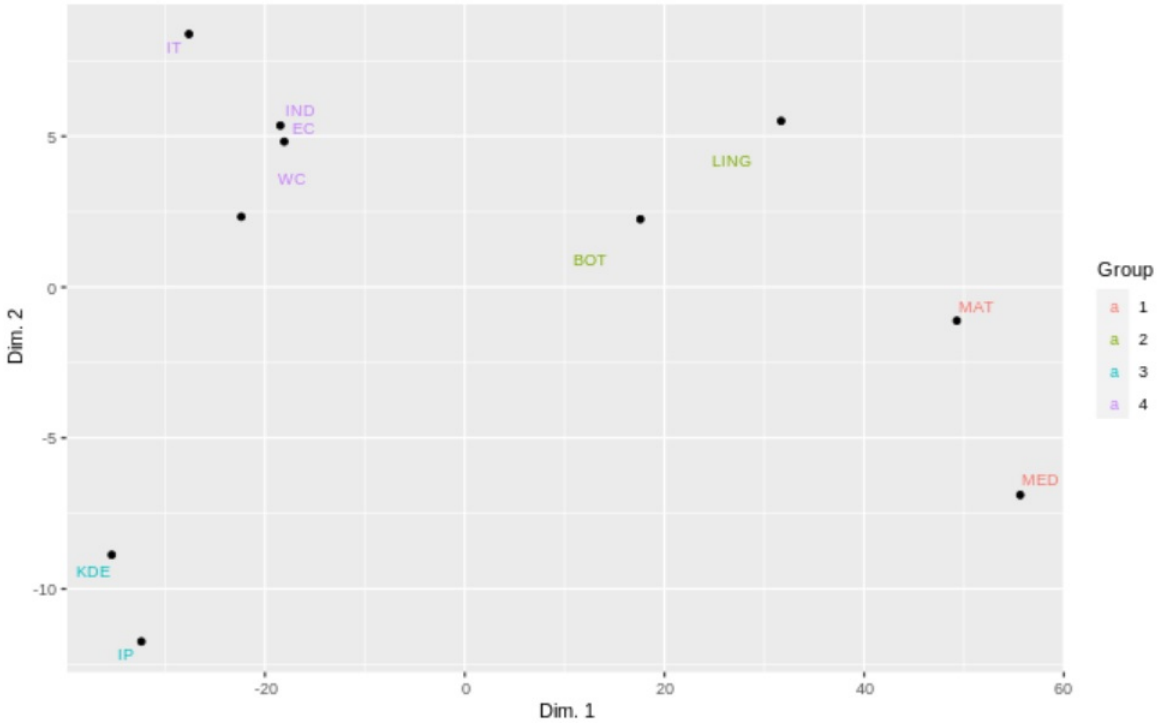
TABLE 5.35: Three features as variables for multidimensional scaling

Feature	BOT	EC	IND	IP	IT	KDE	LING	MAT	MED	WC
Linearity	29	57	57	78	63	79	15	4	2	62
Cyclicity	17	13	12	9	15	13	3	0	1	14
Variation	29	53	54	48	61	51	28	13	4	53

[a] BOT = Botany; EC = Evolutionary computing; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; WC = Wireless computing

Table 5.35 shows the values for cyclicity, linearity and variation that were obtained for the functions which were imported into the dataframe used for MDS. The tailor-made script used is available in Appendix A.13.

The results for the automatically-derived clusters created using MDS followed by k-means hierarchical clustering are plotted in Figure 5.15.

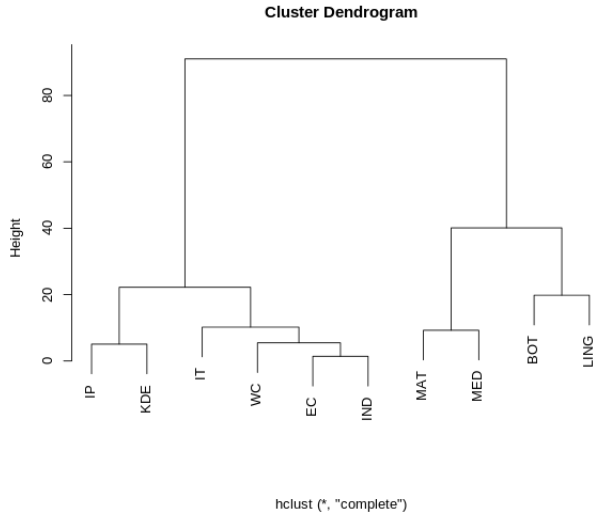


[a] BOT = Botany; EC = Evolutionary computing; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; WC = Wireless computing

FIGURE 5.15: Plot showing the results of the k-means cluster analysis

Figure 5.15 divides the disciplines into what appears to be four clusters with the name of each cluster coloured according to the cluster. As with the manually created *Borromean Rings* framework in Figure 5.14, BOT and LING form one cluster and MAT and MED form a separate cluster. However, in the k-means cluster the remaining six disciplines are further subdivided into two clusters, one of which comprises KDE and IP while the final cluster comprises EC, IND, IT and WC.

The results of the categorization algorithm can be visualized as a dendrogram as shown in Figure 5.16.



[a] BOT = Botany; EC = Evolutionary computing; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; WC = Wireless computing

FIGURE 5.16: Dengrogram visualization

The classification into three broad groups shown in Figure 5.16 matches the grouping made manually using the Borremean rings framework. The similarity between the manual and MDS plots adds evidence in support of the *Borromean Rings* framework. The first branch of the cluster dendrogram comprises the six disciplines that based on their cyclicity and variation are categorized into the same segment in the *Borromean Rings* framework. The second branch splits into the two other clusters namely the linear cluster of MAT and MED and the linear-variation cluster of BOT and LING.

5.7.5 Sub-question 5 conclusion

To sum up, the similarities in rhetorical organization between the disciplines can be seen using the dimensions of linearity, cyclicity and variation. Disciplines that report one primary outcome, such as the success of a trial or result of an experiment, tend to be linear. This includes disciplines such as medicine and botany.

Disciplines which are more homogeneous with greater degrees of shared knowledge lack the necessity for lengthy introductions to frame the importance of the research area.

Disciplines which focus on the development and evaluation of artefacts tend to harness the cyclicity dimension. The resultant abstracts are, therefore, non-linear. For example, disciplines such as industrial electronics and wireless communications develop systems and then run simulations. The method and results of both the development and evaluation phases are reported.

The results of the hierarchical clustering add validity to the *Borromean Rings* framework as the same clusters were achieved. The hierarchical clustering achieved slightly more fine grained results for the set of engineering and information science abstracts (EC, KDE, IND, IP, IT and WC). Specifically, KDE and IP were more closely aligned. A close inspection of the most frequent permutations in each of these disciplines confirms their greater similarity. A close comparison of their most frequent adjacent pairs of rhetorical moves shows that in both domains the three most common adjacent pairs are identical.

5.8 Sub-question 6: Differences in rhetorical organization

5.8.1 Preamble

This subsection presents the answer to the sixth sub-research question namely:

> "What are the differences in rhetorical organization between the disciplines?"

As with Sub-question 5 which focused on similarities and was discussed in Subsection 5.7, this question is also exploratory and so no hypotheses are tested.

When disciplines are grouped by similarity manually or using a clustering algorithm, disciplines that were more similar were grouped together. This means disciplines are assigned to a *proximate genus*, which is the class with specific defining elements. Disciplines can be more finely grained within a *proximate genus* through the identification of a *specific difference*. Axiomatically, disciplines that did not display similarities therefore displayed differences.

Three types of differences were discovered that relate to type of abstracts, namely:

1. whether the abstracts are traditional or not
2. three distinctive dimensions of linearity, cyclicity and variation; and
3. the disciplinary focus on particular moves.

Each of these differences are explored in the following subsections.

5.8.2 Traditional vs. Non-traditional abstracts

Traditional abstracts are short text-based summaries presented in the form of prose written as a single or multiple paragraphs. Non-traditional abstracts are those that deviate from this style.

Both MED and MAT abstracts were non-traditional, resulting in differences that manifested in the rhetorical organization. Both these disciplines eschewed traditional abstracts in favour of two modern innovations. Medical abstracts used structured abstracts. Materials science abstracts used graphical abstracts.

Structured abstracts

Structured abstracts affected the linearity dimension by prescribing the order of reporting moves. Headings are provided to help readers identify the relevant section with the abstract. Medical structured abstracts were comparatively long at approximately twice the mean length for the corpus. This may be due to their standalone nature. Abstracts have been argued to be both a genre and a sub-genre dependent on whether the abstract is viewed as a text that is to be read in conjunction with its accompanying article or not. Specialist informants in the medical profession asserted that they expected practicing medical doctors to only refer to the abstracts and, in most cases, there would be no need for doctors to consult the full article.

Huckin (2001) found that the PURPOSE MOVE was rarely used in biomedical articles. However, given the stipulation that a PURPOSE MOVE must be used in the British Medical Journal, all the research abstracts in the disciplinary sub-corpus for medicine contained the PURPOSE MOVE.

Graphical abstracts

Graphical abstracts were limited to up to 200 words and an image file. The ability to package information in image enables authors to convey information that is difficult to express in words, such as molecular structures, graphs, charts and photographs. The majority of images were found to show or represent the results of the research and so in a multimodal analysis would be classified as RESULT MOVE.

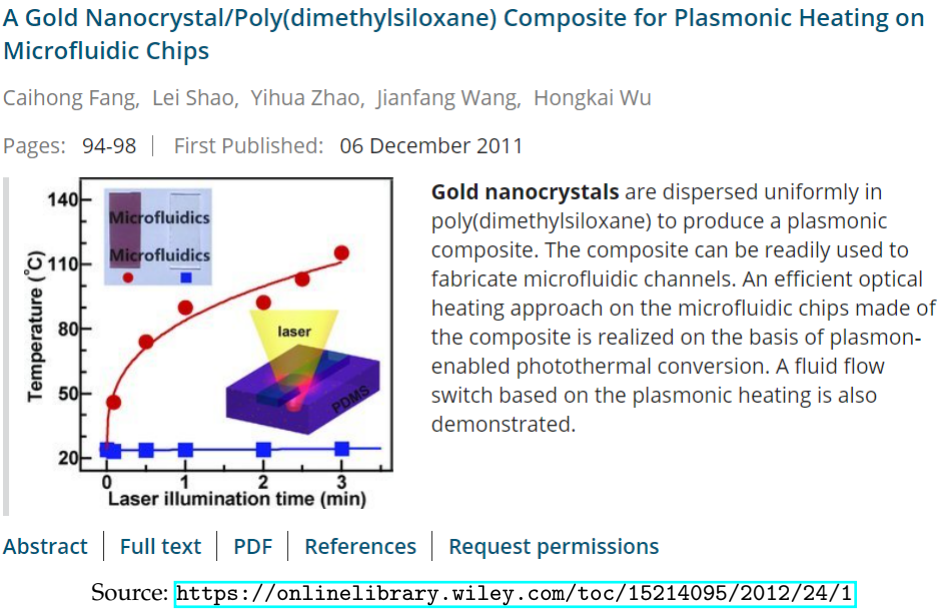


FIGURE 5.17: Example of graphical abstract from the journal *Advanced Materials*

Axiomatically, the use of a visual is the most obvious difference. However, the pithy nature of the short abstracts also affected the length of the move permutations.

Since, MAT abstracts were significantly shorter, there was less likelihood that abstracts contained multiple moves. In fact, 41% of the move permutations for the MAT abstracts comprised a single RESULT MOVE, a further 39% comprised two moves, namely RD (20%), MR (15%) or IR (4%). The only frequent three-move permutation was MRD with five instances, viz. a 5% share of the corpus. The remaining permutations had frequencies of less than four.

5.8.3 Distinctive dimensions

Differences in rhetorical organization can also be shown using the three distinctive dimensions of linearity, cyclicity and variation.

Linearity dimension

In the early days of research abstracts in the middle of the twentieth century, many abstracts acted as a summary and followed a structure that closely resembled the four-move IMRD structure. However, with increased specialisation and to cope with the disciplinary variation, the linearity dimension now includes non-linear elements. With a five-move IPMRD structure, linear abstracts are expected to use moves in the same order as prescribed by the sequence. However, six of the ten disciplines selected did not. Those six disciplines cycled through moves meaning that the second move in the first cycle appears before the first move in the second cycle. Disciplines which are linear display less variation since the position of moves is limited sequentially. It is also not possible to be cyclical and linear, since any cyclical structures create non-linear structures.

Cyclicity and variation dimensions

Cyclicity most commonly occurred with the adjacent pair of MR with the most common cyclical pattern being MRMR. Other moves, however, also were used cyclically. For example, IMIM was also used particularly when the authors were providing some just-in-time background knowledge to help the reader understand a concept related to the method. The likelihood of a delayed INTRODUCTION MOVE might be dependent on the expected shared (or unshared) knowledge among the readers that the abstract was intended for. For complex concepts, providing pertinent information just before it is needed reduces the burden on the short-term memory of the reader. This means that rather than providing a three-part INTRODUCTION MOVE followed by a three-part METHOD MOVE cycling through the INTRODUCTION MOVE and METHOD MOVE three times is likely to make the abstract easier to comprehend.

The split between disciplines showing cyclicity and those that do not is dependent on the nature of the research undertaken. Disciplines that created an artefact and then measured particular features of the functionality of the artefact used cyclicity. This includes the development of algorithms, systems and methods and the subsequent measuring of the speed, accuracy or other feature of the artefact.

5.8.4 Disciplinary focus

Abstracts may begin with hook that catches the attention of the reader and focuses the reader on a novel, significant or substantive aspect of the research. This hook tends to contain an INTRODUCTION MOVE, PURPOSE MOVE or RESULT MOVE. Based on this emphasis, the abstract may be described as introduction-focused, purpose-focused or result-focused. Focus can be judged based on the rank frequency of rhetorical moves. This focus varies by discipline, and so brings to light another difference.

Each community of practice values different qualities. By analyzing the most frequently-occurring move at the beginning of the most frequent permutations for each disciplines, the initial-focus for each discipline can be determined.

Table 5.36 shows the initial-focus for the most frequent permutations of rhetorical moves for each of the ten disciplines.

TABLE 5.36: Initial-focus[a] of rhetorical move[b] permutations by discipline[c]

BOT	No	**EC**	No	**IND**	No	**IP**	No	**IT**	No
I	72	I	36	P	24	I	34	I	10
		R	6	I	6	R	15	R	9
		P	5	R	4			P	8
KDE	No.	**LING**	No	**MAT**	No	**MED**	No	**WC**	No
I	43	I	38	R	61	P	97	I	18
		P	33	M	20			P	10
				I	4			R	4

[a] inclusion based on threshold frequency of 4
[b] I = Introduction; P = Purpose; M = Method; R = Result; D = Discussion
[c] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing

Introduction-focused

Of the most common move permutations, those beginning with an introduction-focus were the most frequent in seven disciplines, namely: BOT, EC, IP, IT, KDE, LING and WC. In some of these disciplines, the percentage of abstracts with an introduction-focus was significantly higher. The most overtly introduction-focused disciplines were BOT and KDE with all the most common permutations beginning with an INTRODUCTION MOVE. LING and IP also displayed high level of introduction focus.

One explanation for this is that Unlike in natural sciences where researchers frequently work on the same or very similar problems, researchers in social and applied sciences tend to work on different problems. Thus, there is less of a shared context, which means that abstracts in the social sciences are more likely to need to contextualize the research by familiarizing the reader with the background details of the problem.

Purpose-focused

The MED sub-corpus contained only Purpose-focused abstracts in the most common permutations due to the prescribed structure dictated by the structured abstract format.

Method-focused

In disciplines in which novelty is valued, a researcher who develops a new method is likely to front the METHOD MOVE. However, although method-focused permutations occurred, it was less common. Materials science was the only discipline with method-focused abstracts being among the most frequent permutations. However, the method-focus abstracts (n = 20) were far fewer than result focus (n = 61).

Result-focused

In disciplines in which increases in performance are valued, researchers are more likely to begin by announcing specific details in a RESULT MOVE. Materials science (n = 61) was particularly notable for this will the highest number of frequent permutations commencing with a RESULT MOVE. Result-focused abstracts ranked in the top three for five other disciplines (EC, IND, IP, IT and WC). Yet, despite its widespread usage, the raw frequency was relatively low varying from the threshold value of 4 abstracts to a maximum value of 15 in the IP sub-corpus.

5.8.5 Sub-question 6 conclusion

In summary, the most obvious differences among the research abstracts were related to the overall format. This resulted in two categories of tradition or non-traditional. The rhetorical organisation in non-traditional abstracts was affected by the choice of format. Structured abstracts prescribed headings which correlated almost identically to the rhetorical move used in each of the headed sections. Graphical abstracts used an image to supplement a short text abstract. The brevity of the abstract resulted in rhetorical move permutations that were comparatively shorter, ranging from one to three moves.

Differences were noted in the dimensions of linearity, cyclicity and variation. Abstracts were either linear (e.g. BOT and MED) or non-linear. Abstracts in some disciplines were cyclical (e.g. EC, IND, IP, IT, KDE and WC) while others were non-cyclical (e.g. BOT and MED). As it is not possible to be both linear and cyclical, these subsets were mutually exclusive. The corpus of abstracts showed variation to different degrees. Some corpora were more homogeneous and displayed little variety (e.g. MAT and MED) while the other disciplines displayed a range of different permutations. The disciplines that displayed cyclicity tended to have more variation.

There were also disciplinary differences in the choice of move to commence an abstract. This move could be considered the focus of the abstract. Introduction-focused abstracts were overall the most common, but some disciplines differed. The most prominent differences were MED with almost all abstracts being purpose-focused and IND with a strong tendency for purpose-focused abstracts. MAT had an extremely strong tendency for result-focused abstracts.

5.9 Implications and applications

> *Truth is truth. Implications are subjective. People will hear your words and draw their own conclusions.*
> - Neal Shusterman, American writer

5.9.1 Preamble

This section aims to draw out and discusses the theoretical implications and practical applications that have been stated or alluded to in the description and discussion of the research results for each sub-research question in this chapter.

Subsection 5.9.2 discusses the disjuncture between the advice that novice writers of scientific research abstracts receive and the abstracts that re published in top-tier journals. Subsection 5.9.3 recapitulates the three dimensions along which scientific abstracts may be categorized. Subsection 5.9.4 summarizes the proposed *Borromean Rings* framework stemming from the three dimensions. Unlike IMRD and IPRMD, this framework does not prescribe any order, but classifies disciplines based on the status of the three dimensions.

5.9.2 Prescriptive-descriptive disjuncture

The set of five moves, (INTRODUCTION MOVE, PURPOSE MOVE, METHOD MOVE, RESULT MOVE and DISCUSSION MOVE), covers all the rhetorical moves detected in the corpus. This confirms the suitability of the tag set, which is in line with many previous researchers who used similar tag sets (Hyland, 2004; Lorés, 2004; Pho, 2008; Salager-Meyer, 1992; dos Santos, 1996; Tseng, 2011)

The pedagogic literature tends to advocate the use of either the four-move structure IMRD nor the five-move IPMRD rhetorical structures. However, despite the validation of the move tag set, neither IMRD nor IPMRD adequately describe the actualized move structure in many scientific disciplines. Few abstracts were found to adhere to the IMRD or IPMRD move permutations, which were just two of the 196 move permutations identified.

Thus, a notable theoretical implication stemming from the results of this study is that there is a disjuncture between the prescribed rhetorical organization in the pedagogic literature. This confirms the preliminary result that was based on five scientific disciplines (EC, KDE, IP, IT and WC) and was reported in Blake (2015b).

Researchers focused on genre have often written about moves and sub-moves, but have tended to focus on the presence or absence of the moves rather than the sequence in which the moves occur. This study specifically focused on identifying the sequences and in doing so revealed the discrepancy between what is advocated and what is actualized.

A key take-away from the discovery of the wide variation in rhetorical organization is that teachers who advocate the use of IMRD or IPMRD as the basis to create scientific research abstracts may be doing a disservice to their students. Without knowledge of the specific disciplinary conventions, prescribing a move structure may do more harm than good. Teachers of research writing should avoid using blanket statements, such as "When writing abstracts, follow the IRMD system." Although this advice may be appropriate for some disciplines, in other disciplines, IMRD-style abstracts are rare, especially in WC and IND which make frequent use of cycling through the METHOD MOVES and RESULT MOVES to report the development of an artefact and then report the evaluation of the performance of the artefact. As a starting point, describing the IMRD structure as a general organization principal may help learners create an appropriate schema, but that should not be the final model.

5.9.3 Three dimensions

The IMRD and IPRMD structures do not sufficiently explain the rhetorical organization in this corpus. Thus, an alternative way of viewing the rhetorical organization was needed. The three dimensions of linearity, cyclicity and variation were found to be useful in mapping and categorizing disciplinary corpora.

Hitherto the assumption appears to have been that research abstracts tended to follow a generally linear path starting with a PURPOSE MOVE or INTRODUCTION MOVE and finishing with a DISCUSSION MOVE or CONCLUSION MOVE.

However, in this corpus, non-linear patterns were detected in which a move is fronted. One example of a fronted move is in the two-move permutation RM. Using the optional move explanation may account for the omission of the INTRODUCTION MOVE and DISCUSSION MOVE, but does not explain the reverse order of the RESULT MOVE and METHOD MOVE. The initial move occupies the prime position, and so researchers place moves here that will convince readers of the novelty or significance of their work. The choice of move is dependent on the values of each discourse community.

Another form of non-linear pattern is cycling through pairs or trios of moves. The cycling through METHOD MOVES and RESULT MOVES in research abstracts often occurs when the authors describe the development and evaluation of an artefact. This is particularly common when describing the development of algorithms, architecture or models and their subsequent evaluation. Assuming the abstract begins with an INTRODUCTION MOVE and ends with a DISCUSSION MOVE, the permutation may be represented as IMRMRD.

However, the first METHOD-RESULT cycle represents the development of the process while the second METHOD-RESULT cycle represents the evaluation cycle. Recently, Rau (2019) and Rau and Antink-Meyer (2020) labelled abstracts that cycle through METHOD MOVES and RESULT MOVES as INTRODUCTION, PROCESS, TEST and CONCLUSION (IPRC) where both the PROCESS MOVE and TEST MOVE comprise two sub-moves (i.e. METHOD MOVE and RESULT MOVE). It should be noted that although cycling through METHOD MOVE and RESULT MOVE was common, cycling occurred for other pairs of moves.

The combination of move repetition and non-linear structures resulted in almost 200 move permutations. Most studies on rhetorical organization have been confined to the easier-to-read disciplines of humanities, linguistics and medicine. This may explain why this degree of variation has not been documented in detail by any other researchers.

To help scientific writers understand how moves are used in their disciplines, an awareness-raising activity can be used (Schmidt, 2012). One such activity could be to map the moves from idealized five-move abstract (e.g. IPMRD) to moves found in a published abstract in their discipline.

Figure 5.18 shows a scenario found in the corpus in which the complete set of five moves from the idealized abstract (shaded in gold) are used in an actual abstract, but two of the rhetorical moves are used twice, creating a repeating pattern by cycling through two moves.

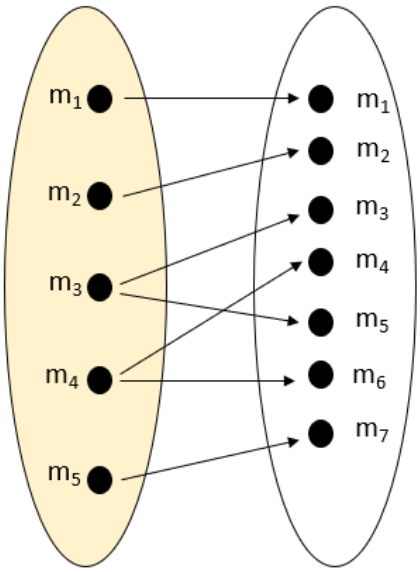


FIGURE 5.18: Bipartite graph mapping idealized abstract to abstract showing cyclicity

Figure 5.19 shows another scenario found in the corpus.

This time only two of the idealized set of rhetorical moves are utilized in the actual abstract. The order of the moves is also reversed with the latter move fronted.

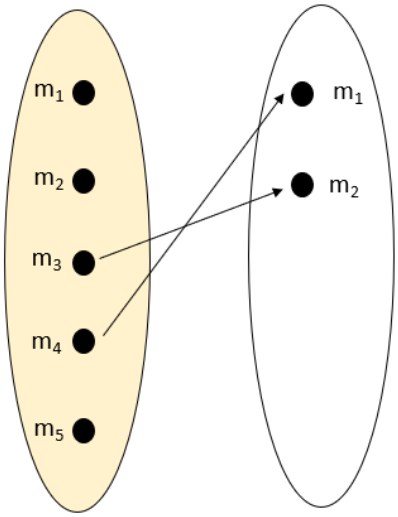


FIGURE 5.19: Non-bipartite graph mapping idealized abstract to abstract showing fronting

This fronting has the effect of adding emphasis to the fronted move. This is analogous to the use of fronting within a sentence to emphasize a particular clause element.

5.9.4 *Borromean Rings* framework

> *All models are wrong, but some are useful.*
> - George E.P. Box, British statistician

Genre theories and other models of language, such as functional grammars (Halliday, 1994) and pedagogic grammars (Murphy, 2012), are reductive (Sarkar, 1992). Knox (2013) asserts reduction is necessary to deal with the complexity and dynamism of language. Models are created for particular purposes and specific audiences. There is currently no model that provides a comprehensive picture of scientific research abstracts in terms of the rhetorical organization of moves or the lexical realization within those moves. Given the reductive nature of models, there will inevitably be some divergence between the idealized model and the reality of the object being modelled (in this case the scientific research abstracts). This divergence can be measured and/or estimated to a particular degree of accuracy. The main aspect, however, is the practical or theoretical use that the model can positively contribute to society.

Previously, researchers have tried to explain deviations from their selected linear rhetorical structures in terms of the omission of moves. The concept of obligatory and optional moves (e.g. Huckin, 2001; Nwogu, 1990; Pho, 2008) were created to explain these differences. The wide variation in rhetorical organization discovered in this study cannot be explained by obligatory and optional moves.

The three dimensions of variation, linearity and cyclicity combine in seven ways, creating seven categories into which abstract types may be classified plus the extra category of zero dimensions, making eight categories. Building on the discovery of

three dimensions, the *Borromean Rings* framework was developed. This theoretical framework can be used by both researchers and teachers of scientific writing.

Writers and teachers of writing alike can use the *Borromean Rings* framework to evaluate whether individual abstracts (e.g. draft abstracts) are similar in terms of rhetorical organization with prototypical abstracts occupying the same region of the *Borromean Rings* framework. By assessing the linearity, cyclicity and variation, writers and teachers gain a better understanding of what their community of practice values. In addition, knowing which disciplines are more similar can help organizers of research writing courses or seminars group students more appropriately. Rather than having to explain the multiple different ways that abstracts can be written, disciplinary specific advice could be given by grouping writers according to categories using the *Borromean Rings* framework.

5.10 Chapter Summary

The compiled corpus of scientific research abstracts was relatively homogeneous, with a mean sentence length of over twenty words and very low readability scores. All the disciplines bar two adopted traditional abstracts. Medical abstracts, however, were structured using prescribed headings. Materials science used graphical abstracts. These differences impacted the rhetorical organization.

Five moves of INTRODUCTION MOVE, PURPOSE MOVE, METHOD MOVE, RESULT MOVE and DISCUSSION MOVE were annotated. All five moves were present in nine disciplines. The INTRODUCTION MOVE, however, was not used in medical abstracts, presumably due to the combination of a high degree of shared knowledge among the readership and the use of structured abstracts.

For the whole corpus, the most frequent moves were RESULT MOVE (approximately 40%), METHOD MOVE (approximately 30%), and INTRODUCTION MOVE (approximately 20%), while the PURPOSE MOVE and DISCUSSION MOVE were the least frequent at less than 10% each.

The RESULT MOVE was most frequent in half of the disciplines (BOT, LING, MAT, IND and WC). The METHOD MOVE was most frequent for IP, IT and MED while the INTRODUCTION MOVE was more frequent in EC and KDE.

Each of the rhetorical moves was found to occur both before and after every other. The two most frequent adjacent pairs were MR and RD which adhere to what would be expected for users of IMRD. However, the third and fourth most frequent adjacent pairs were RM and IR which do not adhere. Many unexpected pairs of adjacent rhetorical moves, such as DISCUSSION MOVE followed by INTRODUCTION MOVE, were discovered. These were usually the result of research studies that involved the reporting and discussion of two sets of results, typically when an artefact is developed and evaluated.

Slightly under 200 permutations of move sequences were identified. Although this number may seem rather large, when compared to the theoretical number of

permutations, this is just a tiny fraction. Results of three scenarios compared actualized and potential permutations confirm that only a very limited number of possible permutations were realized. Patterns of sequencing were found, and despite the unexpected large number of permutations, similarities in the permutation sequences were discovered at both the level of adjacent pairs of moves and at the level of the whole abstract.

Given that the advice in almost all books for novice writers of scientific abstracts advocates using IMRD, there is a prescriptive-descriptive disjuncture between what is being taught and what was discovered in the corpus. Few abstracts in the corpus followed the IRMD sequence precisely. In fact, even when the the PURPOSE MOVE is subsumed into the INTRODUCTION MOVE, the total number of instances of IMRD is only 26 out of 1000 abstracts.

Three dimensions were determined on which similarities and differences among the disciplinary corpora can be described. The first dimension is linearity. Abstracts with rhetorical moves occurring in the expected order of IPMRD are classified as linear. The second dimension is cyclicity. Abstracts in which adjacent pairs or trios of rhetorical moves are repeated show cyclicity. The most frequently repeated adjacent pair was METHOD MOVE followed by RESULT MOVE. The third dimension is variation, which is judged at the level of corpus or dataset than abstract. Disciplines with a large number of different permutations of moves show high variation.

One of the theoretical contributions of this research is the creation of the *Borromean Rings* framework. This framework comprises three interlocking rings creating eight distinct regions on to which the eight different combinations of the three dimensions can be mapped. Any disciplinary corpus of research abstracts can be mapped onto a region based on the three dimensions.

Armed with the knowledge that scientific research abstracts do not necessarily follow the IMRD or IPMRD rhetorical move structures, teachers of scientific research writing can provide more informed advice to their students. Specifically, teachers are advised to avoid blanket generic advice and encourage student writers to explore disciplinary specific corpora to understand how published abstracts convey the novelty, significance, substance and rigour of their research. The IPMRD move structure is a helpful model to introduce rhetorical organization, but the use of fronting moves and cycling through moves should be introduced to students whose disciplines display variation and cyclicity, such as IND and WC.

Writers need to understand the principles for the selection and sequencing of the rhetorical moves rather than adhering to a proscribed sequence of rhetorical moves (unless, of course, the editorial guidelines prescribe the sequence). Given that a central aim of a research abstract is to persuade readers to read the whole article, the strengths of the accompanying research article need to be highlighted. In research publishing, novelty, significance, rigour and substance are the primary concepts on which research is evaluated. Whichever of these concepts is most relevant to the reader is the element more likely to be fronted. However, disciplinary norms need to

be considered, particularly for novice writers.

This study focused on the statistical analysis of rhetorical organization by identifying the permutations of rhetorical moves identified in a corpus. However, to gain further insight into the reasons for particular permutations would need extensive cooperation with authors to understand how the final rhetorical organization came about. This could be achieved by tracking the history of a draft abstract through a version control system linked to a LaTeX platform, such as *Overleaf*; and then discussing the reasons for the changes in the abstract with the authors.

Chapter 6

Lexical realization

> *You string some letters together, and you make a word. You string some words together, and you make a sentence, then a paragraph, then a chapter. Words have power.*
> - Chloe Neill, American Author. Firespell

6.1 Chapter preview

This chapter describes and discusses the results that can be used to answer the second research question (2.8.3), namely: "What are the lexical features of prototypical moves in abstracts of research articles in the selected scientific disciplines?" The underlying motivation is to understand how lexis is used within and among rhetorical moves and subject disciplines. This knowledge should serve as valuable resource for teachers and learners of scientific research writing.

The main research question is divided into four sub-questions, which are reproduced here:

1. Does the lexical realization differ between the same moves in different disciplines?
2. Does the lexical realization differ between different moves in the same discipline?
3. To what extent does the lexical realization differ between moves?
4. To what extent does the lexical realization differ between disciplines?

As argued the literature review (2.6.2), linguists and language teachers may pigeonhole words into the categories of lexis or grammar; yet grammar is, in fact, realized by lexis. Grammar can, therefore, be viewed grammatically or lexically. This lexis-grammar duality of language is analogical to the wave-particle duality of quantum theory as conveyed by the thought experiment of Schrödinger's cat (Gribbin, 2011). As Halliday (1992, p.64) notes, when a language system is viewed lexically, lexis-like answers are obtained; and, conversely, when the language system is viewed grammatically, grammar-like answers are obtained. Lexical realization in this chapter

is used in the broad sense to refer to both lexis *per se* and grammatical concepts realized by lexis. (Halliday, 1992; Stubbs, 1996). In this study, lexical realization is investigated by "interrogating" the corpus lexically using keyness as the proxy and grammatically using grammatical tenses as the proxy. Here tense does not refer to the past-present dichotomy but refers to grammatical tenses, which are defined as the twelve verb forms of tense and aspect, commonly taught in EFL textbooks, such as *present perfect progressive* and *future simple*.

Keyness was selected as a proxy rather than words *per se* because of its ability to provide an "aboutness". Key words are identified and a ranked by keyness for all the sentences within each move within each discipline. Key word analysis was conducted three ways, namely by discipline, by move and by both move and discipline. Key word analysis considers the form and not the function of words, and so this snapshot of the lexical realization is taken through a lens that is frequency-focused. Manual inspection of the text in the analysis stage is used to triangulate the conclusions drawn. Key words are not necessarily spread evenly throughout a corpus and so the dispersion of key words also needs to be considered (Gries, 2008). Where appropriate, dispersion plots are used to identify the degree to which key words are spread throughout a corpus. Some key words may be unevenly dispersed and show what Church and Gale (1995) term *burstiness* and Gries (2020, p.99) refers to as *clumpiness*. Following the advice from Gries (2008, p.428), both frequency and dispersion are reported.

Grammatical tense was selected as a proxy for grammar. Grammatical tense is used to describe the twelve verb forms commonly taught in EFL textbooks. Grammatical tense therefore includes modality (for future forms), perfect aspects and progressive aspects. Grammar covers a wide range of features at morphological, lexical, phrasal, clausal, sentential and discoursal levels. At phrase level, grammar can focus on nominal, verbal, adjectival or adverbial phrases. Tense plays a central role in EFL teaching and teaching materials. Tense is also a central concern for learners themselves, possibly due to its complexity; the interaction between meaning, form and function; and the lack of one-to-one mapping to mother tongue. Tense can be considered in numerous ways, two of which are syntactical and semantic. The lack of a unambiguous correspondence between the morphology of tenses and the meaning ascribed to the tense complicates any analysis. Tense analysis in this research adopts a syntactic approach, which considers the form of the string (sequence of words) containing a finite verb phrase, and not the underlying grammatical meaning in context of the grammatical tense (Blake, 2020a; Blake, 2020c).

Multidimensional scaling combined with a k-means clustering algorithm was utilized to identify the degree of similarity between moves and disciplines in terms of keyness and tense. Axiomatically, features that do not show similarity are dissimilar and so cluster analysis can also be used to discover differences.

The following four sections describe and discuss the results related to each of the four sub-questions. When combined, the second research question can be answered.

Section 6.2 compares the lexical realization in the sames moves in different disciplines. The results on disciplinary variation that is discovered are analyzed and explained. Section 6.3 analyzes the lexical realization in different moves within the same disciplines. This section shows the effect that rhetorical move has on lexical realization. Section 6.4 extends the analysis of disciplinary variation in moves by investigating the extent to which lexical realization differs by discipline, while Section 6.5 considers the extent to which variation occurs in rhetorical moves within a discipline. Section 6.6 extends the discussion of the results on lexical realization focussing specifically on the theoretical implications and practical applications. The chapter summary 6.7 draws together the key findings.

In this chapter, the INTRODUCTION MOVE, which is absent from the medical abstract corpus is not included in the hypothesis testing. If it were, the hypotheses would not need further investigation as the lack of a move would result in a lexical difference created by the absence of text. The lexical realization will, however, be investigated for the remaining 49 out of 50 sub-corpora (i.e. 10 disciplines x 5 rhetorical moves).

Sub-research questions 7 (Section 6.2) and 8 (Section 6.3) are hypothesis-testing while Sub-research questions 9 (Section 6.4) and 10 (Section 6.5) are exploratory or hypothesis-generating.

6.1.1 Tool selection

> *You want to find the tool for the job, instead of holding a hammer and looking for a nail.*
>
> - George Oster, American mathematical biologist

Table 6.1 provides an overview of the tools used to investigate the lexical realization within the corpus and to find the answers to the research question and sub-questions.

TABLE 6.1: Tools utilized to investigate lexical realization

Purpose	Function within tool	Tool
To rank words by keyness	Keyness-calculator	Open-source Python program
To identify grammatical tense	Tense identifier	Tailor-made Python program[a]
To group similar items	Cluster analysis[b]	Tailor-made R program
To examine word usage	Concordance (KWIC)	AntConc[c]
To rank words by frequency	Word list	AntConc[c]
To plot dispersion	Concordance plot	AntConc[c]
To find co-occuring terms	Clusters/N-grams fn	AntConc[c]
To rank words by keyness	Keyword list function	AntConc[c]

[a] This script improves on the prototypes Blake (2020a), Blake (2020b), and Blake (2020c)
[b] using *K*-means flat clustering algorithm that minimizes Euclidean distance
[c] AntConc (Anthony, 2019) is a popular concordancer to investigate corpora evidenced by over 2,100 citations on Google Scholar

Keyness was investigated systematically using Keyness-calculator, an open-source Python script (See Appendix A.9 for the script). This program incorporates Ratio, the

simple math formula created by Kilgarriff (2009). This open-source script obviates the need to use Sketch Engine (Kilgarriff et al., 2014) to access Ratio, thereby improving the workflow.

The twelve grammatical tenses were identified and counted using a tailor-made Python program created specifically for this project (See Appendix A.11 and Appendix A.12 for the relevant scripts). Earlier versions of this script have been incorporated in three prototype online tense identification tools (Blake, 2020a; Blake, 2020b; Blake, 2020c).

A hierarchical *k*-means clustering algorithm was created by stacking standard functions and adapting the code to create a tailor-made Python script (See Appendix A.8 for the script).

AntConc (Anthony, 2019) was used to investigate the annotated corpus for numerous standard corpus queries, such as examining words in context, plotting dispersion, arranging words by frequency and ranking key words by keyness.

6.2 Sub-question 7: Discipline-specific lexical realization by move

6.2.1 Preamble

This section investigates the first sub-question of the second main research question using the corpus described in 5.2. As there were six sub-research questions in the first main research question, this sub-question becomes the seventh sub-research question. The question is given below.

> "Does the lexical realization differ between the same moves in different disciplines?"

The null and alternative hypotheses are formulated as follows:

> $\mathbf{H}_0$: The lexical realization for each move is similar across all ten disciplines.
>
> $\mathbf{H}_\mathrm{A}$: The lexical realization for each move is *not* similar across all ten disciplines.

These hypotheses can be expressed algebraically as:

$$H_0 : \mathbb{LR}|\mathbb{D} \approx C \qquad \text{for } \{\mathrm{I, P, M, R, D}\}$$
$$H_a : \mathbb{LR}|\mathbb{D} \not\approx C \qquad \text{for } \{\mathrm{I, P, M, R, D}\}$$

where $\mathbb{LR}$ is the lexical realization, $\mathbb{D}$ represents the set of disciplines, C represents a constant value and *I, P, M, R, D* are the five rhetorical moves.

The term *similar* is operationalized as to within a range of 20%, but ignoring obvious outliers that may otherwise skew the results.

Lexical realization is first considered using word frequency. This is followed by a more in-depth analysis using keyness and grammatical tense.

6.2.2 Word frequency

Investigating word frequency provides a concrete picture of the lexical landscape. As Biber (2004, p.176) notes, frequency data extracted from a corpus helps identify particular patterns of usage that may have otherwise remained hidden. Almost all of the most frequent words are grammatical or function words, such as: *the, be, to, of* and *and*. This means that insights into disciplinary-specific or move-specific lexical realizations are less easy to perceive when considering word frequency alone. When the polsemy of just the five words (*the, be, to, of* and *and*) is considered, these words account for approximately 80 different meanings according to Wikipedia[1].

TABLE 6.2: Top ten most frequent words by discipline[a]

Rank	BOT	EC	IND	IP	IT	KDE	LING	MAT	MED	WC
1	the	the	the	the	the	the	the	the	of	the
2	of	of	of	of	of	of	of	of	the	of
3	in	a	and	a	and	and	and	and	and	and
4	and	and	a	and	a	a	in	a	in	to
5	a	to	to	to	is	to	to	in	to	in
6	to	in	is	in	in	in	a	is	with	a
7	that	is	in	is	to	data	that	to	for	is
8	is	this	for	image	for	is	on	with	a	we
9	by	for	are	for	that	we	this	for	was	for
10	for	that	this	we	are	for	for	by	were	this

[a] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing
[b] According to AntConc 3.5.9 (Anthony, 2019)

It is unsurprising that the most frequent words tend to be function or grammatical words. This is because vocabulary follows a Zipfian distribution (Bochkarev, Lerner, and Shevlyakova, 2014). Table 6.2 shows the top ten words in each discipline by frequency of occurrence as ranked by AntConc 3.5.9 (Anthony, 2019). This list is dominated by grammatical words. However, of interest are a personal pronoun *we* and two content words *image, data* which were sufficiently disproportionally frequent to be ranked in the top ten.

Of the hundred words, the nominative case of the first-person plural personal pronoun, *we*, occurs three times. Authorial choice of pronoun has been researched extensively by numerous researchers. Bakhtin (1986) in one of the earliest analyses of this particular discoursal choice asserts that *I*-authors position themselves as single authors and as such are independently responsible for the content of their research while *we*-authors position themselves as a team of authors who are collaboratively responsible. In order to avoid personalizing the discourse, authors may position themselves as *we*-authors even when they are the sole authors. Vassileva (2002, p.261)

[1] https://en.wikipedia.org/wiki/Most_common_words_in_English

notes that Bulgarian researchers tend to eschew using *I* to create a more "objective" style. Harwood (2005, p.1207) investigated the use of authorial pronouns, *I* and *we*, using a corpus approach and found that "the 'author-evacuated' articles in the hard sciences can be seen to carry a self promotional flavour with the help of personal pronouns."

The two content words, *image* and *data*, are both central to each research domain and occur in multiple compound nouns. Compound nouns are one of the strongest examples of collocation, i.e. the "pattern of word occurrence in texts" (Sinclair, 1991, p.18). Table 6.3 shows the most frequent compound nouns (bigrams) for *data* in the KDE corpus. Their prevalence may help explain the high number of occurrences of the term *data*.

TABLE 6.3: Compound nouns formed with *data* in the KDE[a] corpus[b]

Rank	Frequency	Range	Bigrams
1	56	5	data sets
2	17	4	data set
3	9	4	data mining
4	7	3	data points
5	7	3	data structure
6	6	3	data objects
7	6	3	data streams
8	5	3	data items
9	4	2	data cubes
10	4	2	data sources

[a] KDE = Knowledge, data and engineering
[b] Identified using Cluster/N-gram function in AntConc 3.5.9

In a similar vein, Table 6.4 shows the most frequent compound nouns (bigrams) for *image* in the IP corpus. Likewise, these collocations may help explain the frequent usage of the word *image*.

TABLE 6.4: Compound nouns formed with *image* in the IP[a] corpus[b]

Rank	Frequency	Range	Bigrams
1	13	3	image registration
2	10	4	image processing
3	9	3	image quality
4	7	4	image data
5	6	2	image luminance
6	6	3	image restoration
7	5	3	image segmentation
8	4	1	image patches
9	4	2	image refocusing
10	4	3	image sequence

[a] IP = Image processing
[b] Identified using Cluster/N-gram function in AntConc 3.5.9

As Tables 6.3 and 6.4 reveal, technical terminology can be discovered through simple frequency searches. However, in order to narrow down the focus to disciplinary-specific and move-specific lexical realizations, a key word approach is adopted, which is described in the Subsection 6.2.3.

6.2.3 Keyness

Words and magic are two powerful forces that can change the world.
- Amy Neftzger, The War of the Worlds

Word clouds

Analysis of key words shows which words are used disproportionately frequently and highlights the lexical sets that the community of practice values and, no doubt, expects members to use. Key words provide a good indication of "aboutness" (Phillips, 1989). (ibid.). However, the process of comparing the relative frequencies of lexical items only considers the frequency of the form of the word tokens, and not their meaning or function. Despite this limitation, keyness is the best available tool to get a picture of what a text is about. Word clouds may be used to visualize the proportional frequency of words in a text as shown in Figures 6.1 and 6.2.

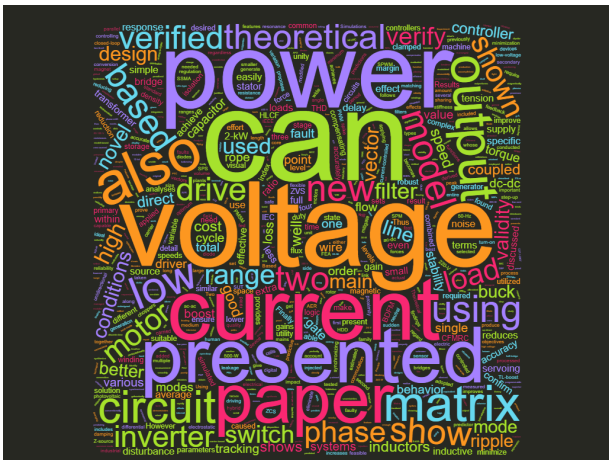


FIGURE 6.1: Word cloud showing vocabulary in the RESULT MOVE of Industrial electronic [IND] abstracts

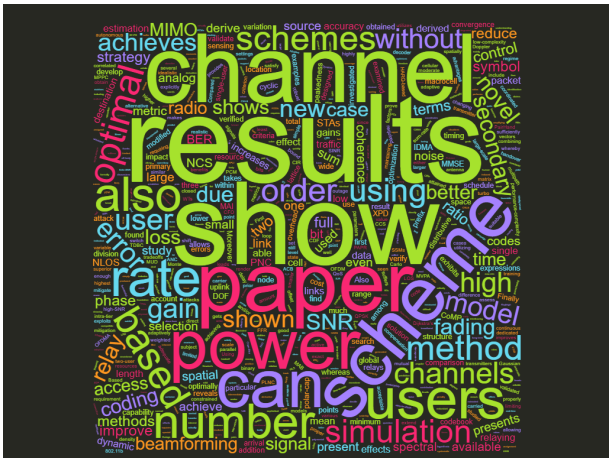


FIGURE 6.2: Word cloud showing vocabulary in the RESULT MOVE of Wireless communications [WC] abstracts

Word cloud visualization is achieved by making the size of the font of a word proportional to the relative frequency of content words within a text. A stop word list of grammatical words was employed to focus the word cloud on the text-specific words rather than the typical frequency distribution of words. Word cloud data visualization does not compare the vocabulary within a text to any benchmark or reference, but simply compares the frequency of a particular lexical item to the sum of the lexical items. Although word clouds can provide a holistic overview, particularly when a limited number of words dominate a corpus, it is difficult to compare word clouds objectively. Thus, rather than using raw word frequency or word clouds, Key words ranked by keyness is used as the proxy to view the vocabulary-side or lexis-focus of lexical realization.

Top ten key words for the whole corpus

Table 6.5 lists the key words for the corpus as whole using the well-established Brown corpus (W. N. Francis, 1965) as the reference corpus. Scott (2009) asserts that no reference corpus is a bad corpus. Given that the purpose of a reference corpus is to establish a baseline against which comparisons can be made, the Brown corpus has been fulfilling this role for decades.

TABLE 6.5: Top ten key words for whole corpus[a]

Rank	Frequency	Keyness	Effect	Key word
1	692	+ 1946.67	0.0075	based
2	710	+ 1901.67	0.0077	paper
3	610	+ 1817.35	0.0066	proposed
4	634	+ 1596.91	0.0069	data
5	600	+ 1565.69	0.0065	results
6	413	+ 1544.42	0.0045	algorithm
7	433	+ 1030.12	0.0047	using
8	272	+ 1026.01	0.003	algorithms
9	394	+ 959.53	0.0043	performance
10	281	+ 957.74	0.0031	propose

[b] According to AntConc 3.5.9 (Anthony, 2019) using log-likelihood (4-term), $p < 0.05$ with the lowercase word list of the Brown Corpus as the reference corpus

If the lexical realization among disciplines is similar, the key words would be expected to be evenly dispersed. Some key words may be regularly dispersed evenly among abstracts, moves and/or disciplines. Other key words may occur in bursts (Katz, 1996), that is key words occur in close proximity of each other. This irregular distribution means that, at times, the key word is not representative of the majority of the corpus, but an outlier that may skew the result. For this reason dispersion plots are used to visually confirm the distribution of target lexical items throughout a corpus, sub-corpus or individual text.

Colligation and collocation example

To provide a specific example, Figure 6.3 shows the dispersion plot for the key word *propose*, which is ranked tenth overall with a keyness score of +957.74 as shown in Table 6.5. The plot in Figure 6.3 shows the position of each instance of *propose* in the METHOD MOVE with a line vertical black line marking the location. To discover whether this word is uniformly spread, we can inspect the dispersion plot. Inspection shows that only eight of the ten disciplines are represented. There were no hits whatsoever in the disciplines of MAT and MED. Moreover, the disciplines of EC and KDE show numerous occurrences of *propose* as evidenced by multiple thick black lines while BOT and LING show only sporadic usage. The usage of *propose* in the METHOD MOVE shows a much higher degree of colligation in the engineering and information science disciplines of EC, IND, IP, IT, KDE and WC. Thus, the distribution of the key word *propose* is uneven.

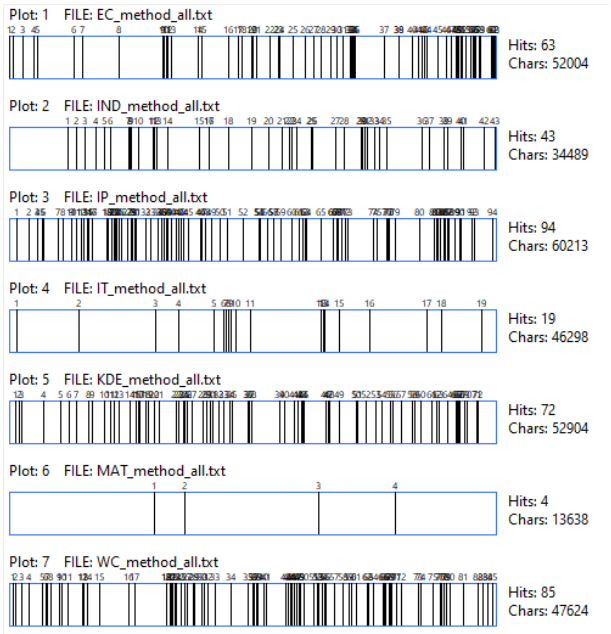


FIGURE 6.3: Dispersion plot of the word token *propose* in the METHOD MOVE
[a] Plotted by AntConc 3.5.9

To get a clearer picture of how *propose* is used, its collocations were investigated. Table 6.6 shows the top 100 collocates of the word *propose* in the whole corpus. From this, the highest ranked collocates were function words, which was expected. Move-specific collocates, such as *paper* and *show*, ranked highly.

Discipline-specific collocates make up most of the remaining words. The more highly ranked collocates are relevant to multiple disciplines, e.g. *effectiveness*, *performance* and *problem*. Collocates that are discipline-specific for just one or two domains also appear including *converter* [IND], *voltage* [IND] and *evolutionary* [EC].

TABLE 6.6: Top 100 collocates[a] for *propose* listed in rank order[b] [c]

the	we	this	of	is
paper	in	and	to	that
for	a	show	are	effectiveness
results	our	on	performance	algorithm
method	based	also	with	problem
demonstrate	algorithms	have	by	first
been	an	image	verify	then
proposed	experimental	efficiency	approach	model
data	be	these	recently	methods
converter	using	scheme	new	channel
two	optimization	framework	range	overcome
validate	system	strategy	schemes	power
or	it	images	hybrid	each
control	analysis	voltage	time	search
research	point	matrix	information	finally
detection	complexity	binary	behavior	which
was	thus	three	technique	superiority
study	simulation	sets	scheduling	real
prove	protocol	novel	its	further
examples	evolutionary	domain	design	confirm

[a] in the whole corpus
[b] using collocate function in AntConc 3.5.9 (Anthony, 2019)
[c] The first item in the top row is ranked 1 while the last item in the bottom row is ranked 100

Dispersion of top ten key words

Taking a broader view, we can examine the pattern of usage of the ten highest ranked key words in each discipline by inspecting Figures 6.4, 6.5, 6.6, 6.7 and 6.8

Figure 6.4 shows the dispersion of the top ten key words listed in Table 6.5 within the INTRODUCTION MOVE. For dispersion plots created in AntConc, the association measure log likelihood was selected. All association measures are biased, but the intrinsic bias of log likelihood to words that are highly frequent seemed preferable to a bias to lower frequency words. The log likelihood measure is the default in AntConc, and is one of the most popular association measures. Each thin vertical link represents one instance of a key word. Only the sub-corpora with at least one hit are displayed. This means that INTRODUCTION MOVE of two sub-corpora (MAT and MED) contained none of the top ten key words. The frequency of the key words was also much lower in BOT and LING. The key words were most commonly used in the disciplines EC and KDE. The top ten key words, therefore, differ by discipline even for the same rhetorical move.

Figure 6.5 shows the dispersion plot for the top ten key words within the PURPOSE MOVE. All ten disciplines are represented, but the frequencies vary considerably with fewer instances in the BOT, IP, MAT and MED sub-corpora compared to the other sub-corpora.

Figure 6.6 also shows the same top ten key word dispersion plot but for the METHOD MOVE. The frequency of occurrence is substantially lower in the BOT, LING and MAT sub-corpora. However, occurrences in the remaining corpora are numerous and dispersed within the remaining sub-corpora.

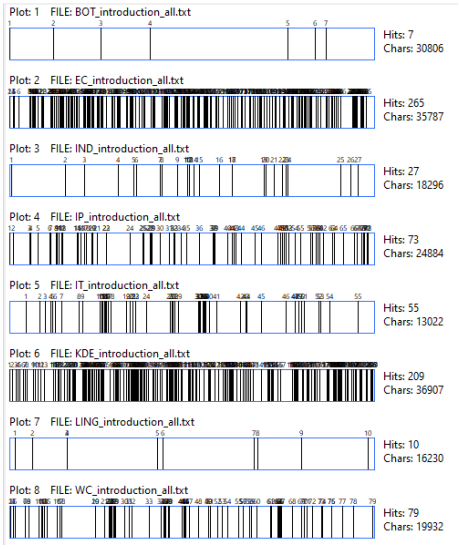


FIGURE 6.4: Dispersion plot of top ten corpus-wide key words in the INTRODUCTION MOVE

[a] Key words calculated using log likelihood in AntConc 3.5.9 with the Brown Corpus word list as the reference corpus

FIGURE 6.5: Dispersion plot of top ten corpus-wide key words in the PURPOSE MOVE

[a] Key words calculated using log likelihood in AntConc 3.5.9 with the Brown Corpus word list as the reference corpus

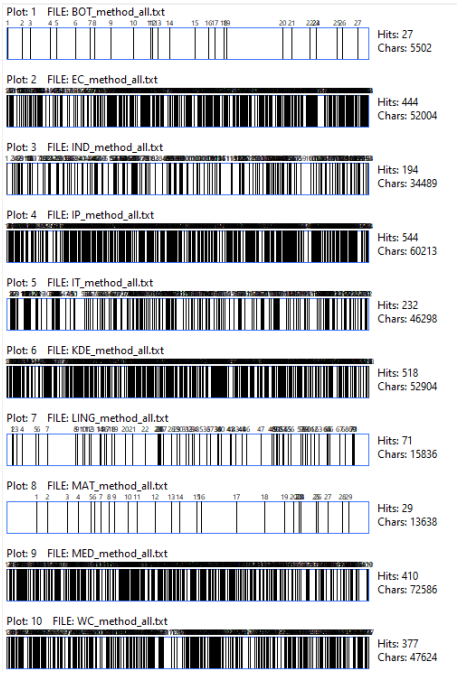


FIGURE 6.6: Dispersion plot of top ten corpus-wide key words in the METHOD MOVE
[a] Key words calculated using log likelihood in AntConc 3.5.9 with the Brown Corpus word list as the reference corpus

The dispersion plot of the same top ten key words in the RESULT MOVE is shown in Figure 6.7. BOT, LING, MAT and MED corpora have multiple results, but they are far less frequent and less dispersed than the remaining six sub-corpora.

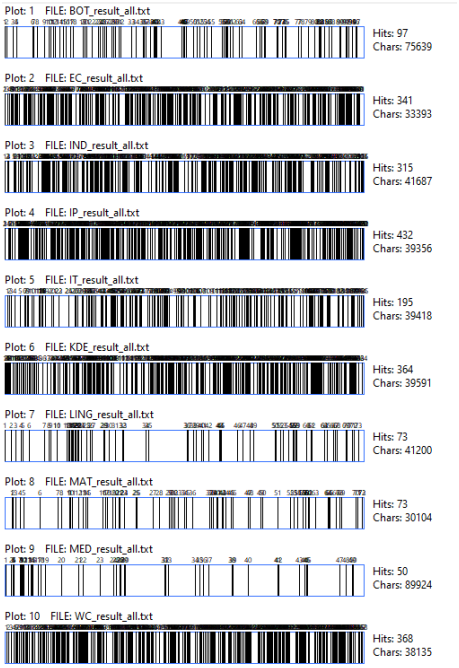


FIGURE 6.7: Dispersion plot of top ten corpus-wide key words in the RESULT MOVE
[a] Key words calculated using log likelihood in AntConc 3.5.9 with the Brown Corpus word list as the reference corpus

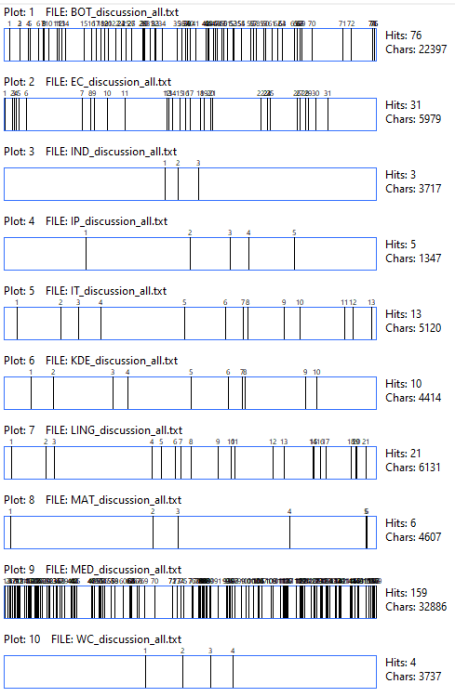


FIGURE 6.8: Dispersion plot of top ten corpus-wide key words in the DISCUSSION MOVE
[a] Key words calculated using log likelihood in AntConc 3.5.9 with the Brown Corpus word list as the reference corpus

Figure 6.8 shows the dispersion plot of the same key words in the DISCUSSION MOVE. There are very few results for most sub-corpora. The BOT and in particular the MED corpora, however, show high frequency and wide dispersion.

Overall, it is clear that disciplinary variation exists for the frequency of occurrence of key words within each move.

Top five key words by move and by discipline

By comparing and contrasting the top five key words within each move in each discipline, we can see whether patterns of usage can be detected. For ease of comparison the results are split into two tables, each showing five disciplines. Table 6.7 shows the top five key words by move and by discipline for the disciplines BOT, EC, IND, IP and IT. Table 6.8 shows the top five key words by move and by discipline for the remaining disciplines of KDE, LING, MAT, MED and WC.

When comparing the key words across each row in Tables 6.7 and 6.8, we can see that most key words are not shared by multiple disciplines. However, a notable exception is the word *paper*, which ranked first in the PURPOSE MOVE in six disciplines (EC, IND, IP, IT, KDE and WC). The key word *results* also ranked in the top five in the RESULT MOVE for four disciplines (EC, IND, LING and WC). Therefore, for two out of five moves there is some degree of shared lexical realization by move.

Through inspection of the concordance lines using the KWIC query function in AntConc, it was discovered that the key word *paper* is often used in the preposition phrase *in this paper*. A regular expression search confirmed this. There were 376

TABLE 6.7: Top five key words[a] by move and by discipline[b] (1 of 2)

Move	BOT	EC	IND	IP	IT
Introduction	arabidopsis	optimization	converter	image	codes
	plants	evolutionary	voltage	images	decoding
	thaliana	algorithms	matrix	interpolation	data
	plant	multiobjective	ac[c]	applications	code
	gene	problems	power	cs[d]	channel
Purpose	arabidopsis	paper	paper	paper	paper
	thaliana	optimization	converter	image	channel
	here	this	dc[e]	video	multiple
	we	algorithm	converters	iterative	channels
	ati[f]	evolutionary	proposed	images	network
Method	arabidopsis	algorithm	converter	image	codes
	thaliana	algorithms	voltage	based	channel
	gene	optimization	proposed	model	algorithm
	mutant	evolutionary	matrix	propose	bound
	genes	based	modulation	algorithm	optimal
Results	mutant	algorithms	proposed	proposed	channel
	protein	algorithm	converter	image	codes
	genes	results	voltage	method	gaussian
	arabidopsis	optimization	results	propose	decoding
	mutants	paper	experimental	based	optimal
Discussion	suggest	algorithm	circuit	image	coding
	role	results	proposed	algorithm	examples
	results	mutation	matrix	video	results
	auxin	algorithms	power	performance	codes
	mediated	evolutionary	suitable	operators	inversion

[a] Sorted by keyness using the keyword function in AntConc 3.5.9 with the Brown Corpus lowercase word list as the reference corpus

[b] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory

[c] ac = alternating current

[d] cs = compressive sensing

[e] dc = direct current

[f] ati = Atg8-interacting [type of protein]

TABLE 6.8: Top five key words[a] by move and by discipline[b] (2 of 2)

Move	KDE	LING	MAT	MED	WC
Introduction	data	media	copolymer	—	wireless
	query	language	uv[c]	—	channel
	queries	studies	detectors	—	multiple
	clustering	research	solar	—	interference
	mining	feedback	electronics	—	network
Purpose	paper	study	polymer	objective	paper
	data	article	acceptor	determine	multiple
	we	this	carbazole	objectives	channel
	problem	investigates	doping	risk	network
	this	internet	ht[d]	assess	channels
Method	data	participants	layer	outcome	channel
	algorithm	language	nanoparticles	participants	we
	query	study	polymer	design	relay
	based	media	ag[e]	setting	propose
	propose	internet	demonstrated	measures	proposed
Results	data	media	polymer	interval	proposed
	paper	results	demonstrated	results	channel
	propose	internet	high	p[f]	results
	based	facebook	organic	confidence	performance
	proposed	findings	graphene	risk	algorithm
Discussion	annotations	implications	applications	conclusions	sampling
	ivat[g]	findings	potential	risk	channel
	preserving	suggest	materials	conclusion	outage
	detection	these	devices	patients	diversity
	query	disagreement	ofets[h]	associated	coding

[a] Sorted by keyness using the keyword function in AntConc 3.5.9 with the Brown Corpus lowercase word list as the reference corpus
[b] KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing
[c] uv = ultraviolet
[d] ht = hexylthiophene-2,5-diyl [P3 HT is a semiconducting polymer]
[e] ag = silver
[f] p = p-value as in p=0.004
[g] ivat = improved Visual Assessment of Cluster Tendency [a data visualization algorithm]
[h] ofets = Organic field-effect transistors

instances of this prepositional phrase in the corpus. On average, slightly under four out of ten abstracts used *in this paper*; yet inspection of the dispersion plot showed that this expression was not used at all in BOT and MED abstracts and only occurred once in LING abstracts. This expression was used to show a contrast between research in previous papers and the research described in their paper, and was used in each of the rhetorical moves. Of the 376 instances, the phrase was used at the beginning of a sentence in approximately 80% of the instances (n = 312). This fronting of a preposition phrase is a realization of information structure by which information is organized according to information-focus, information-flow and end-weight (Biber, Johansson, et al., 1999; Blake, 2015a).

The following examples illustrate the usage of *in this paper*. Example 13 shows the prepositional phrase *in this paper* used at the end of the clause. According to clause grammar, the expected order of clause elements is subject, verb, object, complement and adverbial. Thus, as *in this paper* is used as an adverbial of position, the clause-final position could be viewed as the default. However, in Example 14, *in this paper* is fronted presumably to emphasize the importance and novelty of the content.

(13) Since the variation of motor electromagnetic torque is related to the voltages that are applied to the motor, by analyzing the relationships between stator flux, torque, and voltages, a PMSM torque predictive control scheme is proposed *in this paper*.
[PURPOSE, IND]

(14) *In this paper*, the notion of efficiently invertible extractors is studied...
[METHOD, IT]

One interesting result was the discovery of *Facebook* as a key word for linguistics in the RESULT MOVE. Its usage was investigated using a KWIC search. Examples 15, 16 and 17 show how *Facebook* is used in the RESULT MOVE in linguistics abstracts. As the largest social networking site, it could be assumed that *Facebook* is commonly used as a resource by linguists. However, close inspection of the dispersion plot showed that the 16 instances of the word token *Facebook* occurred in six abstracts. One abstract, from which Example 17 is extracted, used *Facebook* six times to describe research investigating the effect of parents friending their children on this social network site [LING 077]. In another abstract, the use of *Facebook* as the medium of communication during political protests was the research topic. *Facebook* can, therefore, be considered bursty and not representative of the discipline as a whole.

(15) ...communication using Facebook, phone contact, or face-to-face ...
[LING 040]

(16) The link between overall Facebook use and protest activity was explained.
[LING 044]

(17) ...the parent's presence on Facebook also enhanced the child's closeness ...
[LING 077]

6.2.4 Tense

Colligation (Sinclair, 2004c, p.141), the "co-occurrence of grammatical choices". The colligation under consideration here is that of grammatical tense selection. Grammatical tenses, e.g. *past simple*, may occur more frequently in particular rhetorical moves or subject disciplines. However, at present, the distribution of grammatical tenses both within rhetorical moves in scientific research abstracts and within scientific disciplines is unknown. Some pedagogic books advocate the usage of specific grammatical tenses, but the accuracy of the prescribed choices has yet to be confirmed. Three commonly observed suggestions for novice writers are:

1. Use *present simple* to describe general truths, such as when presenting generalized findings in the DISCUSSION MOVE.
2. Use *past simple* to describe completed actions, such as in the METHOD MOVE.
3. Use *present perfect simple* to describe events or states starting in the past but continuing until the present, such as in the BACKGROUND SUB-MOVE of the INTRODUCTION MOVE.

Table 6.9 shows the distribution of grammatical tenses in the whole corpus of 1000 scientific research abstracts.

TABLE 6.9: Grammatical tenses for the whole corpus[a]

Tense	Count	Percentage[b]
Present simple	4600	68.1%
Past simple	1458	21.6%
Future simple	194	2.9%
Present progressive	21	0.3%
Past progressive	2	0.0%
Future progressive	0	0.0%
Present perfect simple	277	4.1%
Past perfect simple	44	0.7%
Future perfect simple	143	2.1%
Present perfect progressive	5	0.1%
Past perfect progressive	2	0.0%
Future perfect progressive	7	0.1%

[a] Calculated to one decimal place

The most frequent grammatical tenses regardless of move or discipline are *present simple* and *past simple* while the least common are *perfect progressive* forms. The frequency of tenses within the whole corpus follows similar proportions to those reported on spoken English (Biber, Johansson, et al., 1999). Table 6.10 shows a summary of the grammatical tenses, or more specifically tense and aspect, described in their seminal corpus grammar, the *Longman Grammar of Spoken and Written English* (ibid.).

Biber, Johansson, et al. (ibid.) categorize the twelve grammatical tenses that are considered in this study into two groups: tensed and modalized. The modalized

TABLE 6.10: Grammatical tenses for spoken English[a]

Tense[b]	Percentage[c]
Present simple	60%
Past simple	18%
Future simple	12%
Present progressive	3%
Past progressive	1%
Future progressive	1%
Present perfect simple	3%
Past perfect simple	1%
Future perfect simple	1%
Present perfect progressive	0%
Past perfect progressive	0%
Future perfect progressive	0%

[a] This table was created based on results published in Biber, Johansson, et al. (1999, pp.452–502)
[b] Future forms are classed as modalized rather than tensed (Biber, Johansson, et al., 1999)
[c] Values rounded to nearest whole number

verbs are those formed by the use of a modal auxiliary verb, such as *will* or *may*. These modalized verb forms may be used to refer to the future. The sum of the percentages for all simple tenses is 90% for the corpus grammar while the sum for all simple tenses in this scientific research abstract corpus is approximately 93% of which 68.1% of the verb forms are *present simple*, 21.6% are *past simple*, and 2.9% *future simple*.

Both accounts of tense in the respective corpora show Zipfian distributions with *Present simple*, and *past simple* being the most frequent. The *present progressive* forms in both sets of analyses are negligible rounding down to 0%.

Holtz (2011, p.71) found that modal verbs occurred "half as frequently in abstracts (0.33%) as in [research articles] (0.77%)". This may account to some extent for the difference in modalized forms between this study of research abstracts and Biber, Johansson, et al. (1999).

TABLE 6.11: Grammatical tenses by discipline[a]

Tense	BOT	EC	IND	IP	IT	KDE	LING	MAT	MED	WC
Present simple	469	607	506	603	537	641	296	195	182	564
Past simple	273	38	46	30	27	45	197	34	753	15
Future simple	10	31	19	21	9	24	13	5	42	20
Present progressive	1	2	2	2	0	4	4	1	1	4
Past progressive	0	0	0	0	0	0	0	0	2	0
Future progressive	0	0	0	0	0	0	0	0	0	0
Present perfect simple	22	60	36	36	12	55	21	8	10	17
Past perfect simple	8	0	0	0	0	1	1	0	32	2
Future perfect simple	4	10	35	22	14	17	6	10	13	12
Present perfect progressive	0	1	1	2	0	0	0	0	1	0
Past perfect progressive	0	0	0	0	0	0	0	0	2	0
Future perfect progressive	0	0	1	1	0	0	0	1	1	3

[a] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory; KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing

Table 6.11 clearly shows the dominance of simple tenses over all the other grammatical tenses. This holds true for each discipline. The frequency of *present simple* far exceeds all other grammatical tenses in all disciplines apart from MED. In the MED corpus, the most frequent tense is *past simple* which is four times more frequent than *present simple*. This focus on *past simple* may be explained by an emphasis on reporting the actions of the authors in the method and the actual results rather than generalizing the findings. Given that medical abstracts were significantly longer, the description of the method and results far exceeded other disciplines. Another notable difference was the use of verbless sentences in the PURPOSE MOVE. Had the verbs been included, the number of occurrences of *present simple* would have been higher. Perfect progressive forms were rarely used with between zero and three instances discovered in each of the disciplinary corpora.

Although relatively rare Examples 18, 19, 20 and 21 show how *past perfect simple* is used in the METHOD MOVE in medical research abstracts.

(18) Proportion of trials for which results had been reported.
[MED 010]

(19) . . . seven days after biopsy, this proportion had increased to 213/1085 . . .
[MED 016]

(20) . . . had agreed to participate before randomisation.
[MED 026]

(21) . . . who had been prescribed escitalopram, citalopram, sertraline, or . . .
[MED 055]

In three disciplines (KDE, IP and EC), *present perfect simple* was ranked the third most frequent tense overall, which might possibly be due to its expected usage in the INTRODUCTION MOVE when showing a research gap.

6.2.5 Sub-question 7 conclusion

As shown in Subsection 6.2.3 and 6.2.4, the lexical realization measured in terms of keyness and grammatical tenses differed between the same moves in different disciplines.

The simple word clouds in Subsection 6.2.3 showed that the vocabulary used in the RESULT MOVE differed between disciplines. The dispersion of the top ten key words in the same moves also differed between disciplines. This was evidenced through dispersion plots of the top ten key words for each move in Subsection 6.2.3. Tables 6.7 and 6.8 showed that the top five key words in each move differed for each discipline. There were more similarities among rhetorical moves in the same discipline than among the same moves in differ disciplines.

Simple tenses account for almost all (93%) of the grammatical tenses in this corpus. This result is slightly higher than reported for spoken English in Biber, Johansson, et al. (ibid.). When tense usage was compared among the disciplines, there were

many commonalities; yet the spread of tenses and the order of the top ranked tenses varied slightly as described in Subsection 6.2.4.

The lexical realization in terms of keyness and grammatical tenses for each move is not similar across all ten disciplines. Disciplinary variation was also discovered for passive voice, which confirms that grammatical choice is affected by discipline. Thus, the null hypothesis $\mathbf{H}_0$ is rejected and the alternative hypothesis $\mathbf{H}_A$ is accepted.

6.3 Sub-question 8: Move-specific lexical realization by discipline

6.3.1 Preamble

Sub-question 8 is:

> "Does the lexical realization differ between different moves in the same discipline?"

The null and alternative hypotheses are formulated as follows:

> $\mathbf{H}_0$: The lexical realization for all moves within each discipline is similar.
>
> $\mathbf{H}_A$: The lexical realization for all moves within each discipline is *not* similar.

These hypotheses can be expressed algebraically as:

$$H_0 : \mathbb{LR} | \forall m_d \{I, P, M, R, D\} = C \qquad \text{for } \mathbb{D}$$
$$H_a : \mathbb{LR} | \forall m_d \{I, P, M, R, D\} \neq C \qquad \text{for } \mathbb{D}$$

where $\mathbb{LR}$ is the lexical realization, d represents the each disciplines, C represents a constant value and *I, P, M, R, D* are the five rhetorical moves.

The term *similar* is operationalized as to within a range of 20%, but ignoring obvious outliers that may otherwise skew the results.

6.3.2 Keyness

If the lexical realization shown by key words is similar for all moves, not only will the ranked frequency of key words be similar, but the distribution or dispersion of key words should be relatively uniform. Dispersion can be assessed visually using the Concordance plot function in AntConc.

When comparing the disciplines in Table 6.7 and Table 6.8, we can see that more vocabulary is shared within a discipline than within a move. Each column shares the same trait, namely that disciplinary-specific vocabulary is used in more than one rhetorical move. Lexical choice appears to be more discipline-specific than move-specific.

Taking the first column for BOT as an example, the dominance of *arabidopsis thaliana* (also known as *mouse cress*) is obvious. On first examination, it seems rather odd that a single plant name could be a key word. In fact, approximately 374,000 is given as the accepted number of plants species in 2016 (Christenhusz and Byng, 2016, p.201) with approximately 2,000 new plant species being added each year (ibid., p.201). Had one of the journals in the corpus been a special issue dedicated to *arabidopsis thaliana*, that would explain its inclusion. However, none of the issues in the sub-corpus were special issues. Further investigation revealed that *arabidopsis thaliana* is the "best-studied model organism in plant biology" (Leonelli, 2007, p.34) and serves as the model plant for genome analysis Meinke et al., 1998, p.678. Thus, despite initial reservations, the key word list appears to provide a sufficiently accurate representation of the "aboutness" of the BOT sub-corpus.

In the second column for EC, the key words *algorithm*, *algorithms* and *optimization* stand out.

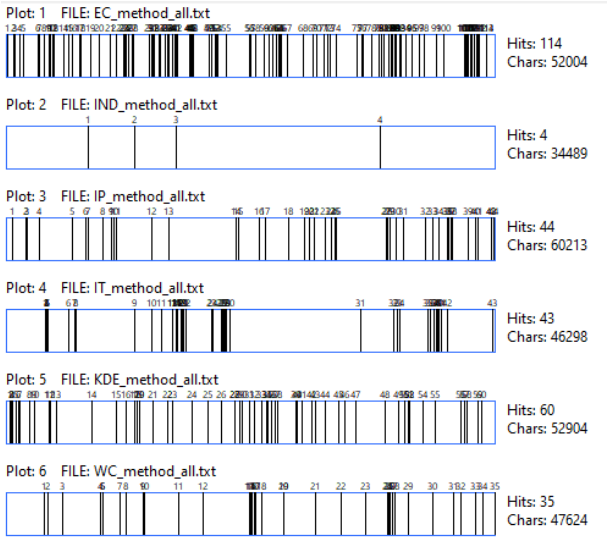
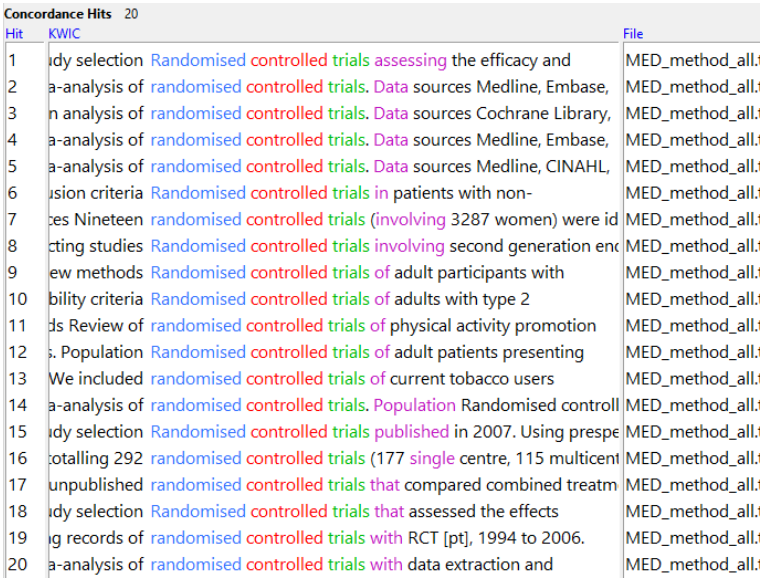


FIGURE 6.9: Dispersion plot of token *algorithm* in METHOD MOVE
[a] Plotted by AntConc 3.5.9

Figure 6.9 shows the dispersion plot for the token *algorithm* in the METHOD MOVE. This shows that this term is used in six of the ten disciplines with *algorithm* being used in 114 instances in EC, which is almost double the number of instances in the KDE sub-corpus. The token *algorithm* was used between 4 and 44 instances in the disciplines IND, IP, IT and WC.

Figure 6.10 shows the KWIC concordance results for *randomised controlled trials* in the METHOD MOVE. All 20 hits were from the medical corpus. This shows that disciplinary variation greatly affects lexical choice. It would be impossible for *randomised controlled trials* to be used in disciplines that do not deal with human or animal subjects, and so in the whole corpus this term could only be expected to occur in LING or MED.



Concordance Hits 20

Hit	KWIC	File
1	dy selection Randomised controlled trials assessing the efficacy and	MED_method_all.t
2	a-analysis of randomised controlled trials. Data sources Medline, Embase,	MED_method_all.t
3	n analysis of randomised controlled trials. Data sources Cochrane Library,	MED_method_all.t
4	a-analysis of randomised controlled trials. Data sources Medline, Embase,	MED_method_all.t
5	a-analysis of randomised controlled trials. Data sources Medline, CINAHL,	MED_method_all.t
6	sion criteria Randomised controlled trials in patients with non-	MED_method_all.t
7	es Nineteen randomised controlled trials (involving 3287 women) were id	MED_method_all.t
8	cting studies Randomised controlled trials involving second generation enc	MED_method_all.t
9	ew methods Randomised controlled trials of adult participants with	MED_method_all.t
10	bility criteria Randomised controlled trials of adults with type 2	MED_method_all.t
11	ds Review of randomised controlled trials of physical activity promotion	MED_method_all.t
12	. Population Randomised controlled trials of adult patients presenting	MED_method_all.t
13	We included randomised controlled trials of current tobacco users	MED_method_all.t
14	a-analysis of randomised controlled trials. Population Randomised controll	MED_method_all.t
15	dy selection Randomised controlled trials published in 2007. Using prespe	MED_method_all.t
16	otalling 292 randomised controlled trials (177 single centre, 115 multicent	MED_method_all.t
17	unpublished randomised controlled trials that compared combined treatm	MED_method_all.t
18	dy selection Randomised controlled trials that assessed the effects	MED_method_all.t
19	g records of randomised controlled trials with RCT [pt], 1994 to 2006.	MED_method_all.t
20	a-analysis of randomised controlled trials with data extraction and	MED_method_all.t

FIGURE 6.10: KWIC concordance for *randomised controlled trials* in METHOD MOVE
[a] Plotted by AntConc 3.5.9

6.3.3 Tense

As can be seen in Table 6.12, the distribution of grammatical tenses varies by rhetorical move.

TABLE 6.12: Grammatical tenses by rhetorical move

Tense	Introduction	Purpose	Method	Result	Discussion
Present simple	921	403	1228	1684	364
Past simple	68	43	401	867	79
Future simple	63	15	20	37	59
Present progressive	15	0	0	4	2
Past progressive	0	0	1	0	1
Future progressive	0	0	0	0	0
Present perfect simple	169	6	41	52	9
Past perfect simple	2	1	23	18	0
Future perfect simple	44	2	34	43	20
Present perfect progressive	3	0	0	2	0
Past perfect progressive	0	0	1	1	0
Future perfect progressive	0	0	3	1	3

However, the ranking of the frequency of tenses follows a similar pattern when grammatical tenses are considered by move with *present simple* and *past simple* being the most frequent tenses overall. However, *present perfect simple* ranks second in the INTRODUCTION MOVE and third in the METHOD MOVE and RESULT MOVE.

Examples 22, 23, 24 and 25 show how *past simple* is used in the METHOD MOVE. This usage follows typical guidelines on tense usage, namely to use *past simple* to describe complete actions.

(22) ...we cloned the restorer gene Rf5 for Hong-Lian CMS in rice ...

(23) ...we utilized dual multiscale Graylevel morphological ...

(24) ...we examined whether our behavioral-learning model ...

(25) ...we took at random 10 high- and 10 low- ...

TABLE 6.13: Top three grammatical tenses[a] by move and by discipline[b] (1 of 2)

Move	BOT	EC	IND	IP	IT
Introduction	PresSimp PastSimp PresPerfSimp	PresSimp PresPerfSimp FutSimp	PresSimp PresPerfSimp FutSimp	PresSimp PresPerfSimp FutSimp	PresSimp PastSimp PresPerfSimp
Purpose	PresSimp PastSimp —	PresSimp FutSimp PastSimp	PresSimp PastSimp PresPerfSimp	PresSimp FutSimp —	PresSimp PresPerfSimp FutSimp
Method	PastSimp PresSimp PresPerfSimp	PresSimp PastSimp PresPerfSimp	PresSimp PastSimp PresPerfSimp	PresSimp FutPrefSimp PastSimp	PresSimp PastSimp FutPrefSimp
Results	PresSimp PastSimp PresPerfSimp	PresSimp PastSimp PresPerfSimp	PresSimp PastSimp FutPrefSimp	PresSimp PastSimp FutSimp	PresSimp PastSimp FutPrefSimp
Discussion	PresSimp PastSimp FutSimp	PresSimp FutSimp PastSimp	PresSimp FutPerfSimp FutSimp	PresSimp FutPerfSimp PresPerfSimp	PresSimp FutPerfSimp FutSimp

[a] Pres = Present; Fut = Future; Simp = Simple; Perf = Perfect
[b] BOT = Botany; EC = Evolutionary computing; IND = Engineering (Industrial electronics); IP = Image processing; IT = Information theory

TABLE 6.14: Top three grammatical tenses[a] by move and by discipline[b] (2 of 2)

Move	KDE	LING	MAT	MED	WC
Introduction	PresSimp PresPerfSimp FutSimp	PresSimp PresPerfSimp FutSimp	PresSimp — —	— — —	PresSimp PresPerfSimp FutSimp
Purpose	PresSimp FutSimp —	PresSimp PastSimp FutSimp	PresSimp — —	PresSimp PastSimp FutSimp	PresSimp — —
Method	PresSimp PastSimp FutPerfSimp	PastSimp PresSimp FutPerfSimp	PresSimp PastSimp FutPerfSimp	PastSimp PresSimp PastPerfSimp	PresSimp FutPerfSimp FutSimp
Results	PresSimp PastSimp PresPerfSimp	PastSimp PresSimp FutSimp	PresSimp PastSimp FutPerfSimp	PastSimp PresSimp PastPerfSimp	PresSimp PastSimp FutSimp
Discussion	PresSimp FutPerfSimp FutSimp	PresSimp FutSimp PastSimp	PresSimp FutSimp PastSimp	PresSimp PastSimp FutSimp	PresSimp FutPerfProg FutSimp

[a] The 12 grammatical tenses are shown in Table 2.14. Abbreviations used are: Pres = Present; Fut = Future; Simp = Simple; Perf = Perfect; Prog = Progressive
[b] KDE = Knowledge, data and engineering; LING = Linguistics; MAT = Materials science; MED = Medicine; WC = Wireless computing

Table 6.13 and Table 6.14 show the most frequent grammatical tenses in each move in each discipline. These tables enable the spread of the tenses across disciplines and rhetorical moves to be compared and contrasted. Although in the INTRODUCTION MOVE the most frequent grammatical tense is *present simple*, *present perfect simple* ranked second in six disciplines, namely EC, IND, IP, KDE, LING and WC. Rather unexpectedly, *future simple* ranked third in the same six disciplines.

Examples 26, 27, 28 and 29 show how *future simple* is used in the INTRODUCTION MOVE.

(26) We will discuss theoretical properties of ...

(27) The problem we will consider in this paper is binary ...

(28) We expect that it will be easier to ...

(29) The performance of these invariants will be demonstrated through ...

Concordance Hits 108

Hit	KWIC	File
1	cription (MD) coding has been a popular choice for robust	IP_introduction_all.txt
2	sense, passive filters have been a very potent technique for	IND_introduction_all.txt
3	veloped. While there have been applications of the wavelet variance	IP_introduction_all.txt
4	to be important and have been applied to tackle single-objective	EC_introduction_all.txt
5	much attention and has been applied to many scientific and	EC_introduction_all.txt
6	entional approaches have been applied to PM machine design	IND_introduction_all.txt
7	The wavelet variance has been applied to a variety of	IP_introduction_all.txt
8	le-objective function have been around for decades. They are	EC_introduction_all.txt
9	ges (Citrus sinensis) have been associated with cardiovascular health, a	BOT_introduction_all.txt
10	18, 849–860, 1992], has been attributed to a resource "bottleneck,"	LING_introduction_all.txt
11	ing leaf primordia. It has been believed that the SAM contains	BOT_introduction_all.txt
12	theoretical analysis has been conducted to study its convergence	EC_introduction_all.txt
13	theoretical analysis has been conducted to study its convergence	EC_introduction_all.txt
14	mplex regimes. This has been confirmed by numerical studies, e.	EC_introduction_all.txt
15	little attention. There has been considerable recent interest in using	IP_introduction_all.txt
16	evolutionary search have been considered in a number of	EC_introduction_all.txt
17	olutionary principles has been contributed in order to overcome	EC_introduction_all.txt
18	gypt, and elsewhere have been credited in part to the	LING_introduction_all.txt
19	projective clustering has been defined as an extension to	KDE_introduction_all.txt
20	n and bucketization, have been designed for privacy preserving microd	KDE_introduction_all.txt

FIGURE 6.11: KWIC regex concordance for *ha(s | ve) been* in the INTRODUCTION MOVE
[a] Plotted by AntConc 3.5.9

To gain greater insight into how these grammatical tenses are realized, results of a search for the strings *have been* or *has been* in the INTRODUCTION MOVE are shown in Figure 6.11. There are 108 concordance hits with almost every instance being an example of *present perfect simple*.

Examples 30, 31, 32 and 33 show instances using the verb *apply*.

(30) ...have been applied to tackle single-objective ...

(31) ...has been applied to many scientific and ...

(32) ...have been applied to PM machine design ...

(33) ...has been applied to a variety of ...

Examples 34, 35, 36 and 37, show instances in which the verb *propose* was used in *present perfect simple* in the INTRODUCTION MOVE.

(34) ...have been proposed for distributed computer networks ...

(35) ...have been proposed for visual tracking, ...

(36) ...have been proposed for mining useful patterns ...

(37) ...have been proposed for text categorization, ...

Instances of *present perfect progressive* used in the INTRODUCTION MOVE were also found as shown in Examples 38, 39 and 40.

(38) ...has been receiving attention as a robust framework.

(39) ...for such an analysis has been lacking.

(40) ...has been receiving attention ...

Examples 41 and 42 show how *present progressive* is used in with the verb *become* in the INTRODUCTION MOVE.

(41) ...is becoming a very promising candidate to ...

(42) ...are becoming increasingly important to individual users

Although grammatical tense was selected as the proxy through which to view grammatical choice, a simple search for passive voice used with regular verbs in the *method move* was carried out by matching strings that include *is*, *are*, *was* or *were* followed by a word ending in *-ed*. The possibility of some false negative results caused by participle adjectives was considered, but few instances were found. Figure 6.12 shows the dispersion plot which found hits in eight disciplines. Occurrences in six disciplines (BOT, EC, IND, IP, KDE and MAT) were in single digits. Passive voice was more frequently used in linguistics with 21 concordance hits while the medical corpus had four times as many hits (n = 108). Biber (1988, p.104) notes that the presence of passive voice is one of the indicators of expository texts. Disciplinary variation is a key factor at play.

6.3.4 Sub-question 8 conclusion

As shown in 6.3.2 and 6.3.3, the lexical realization measured in terms of keyness and grammatical tenses differed between different moves in the same discipline.

There was more similarity between moves in the same discipline than across the same move in different disciplines. Many terms within the same discipline occurred in the top five most frequent words in each move. However, as shown in Tables 6.7 and 6.8 the rank frequency of the top five key words differed in every move.

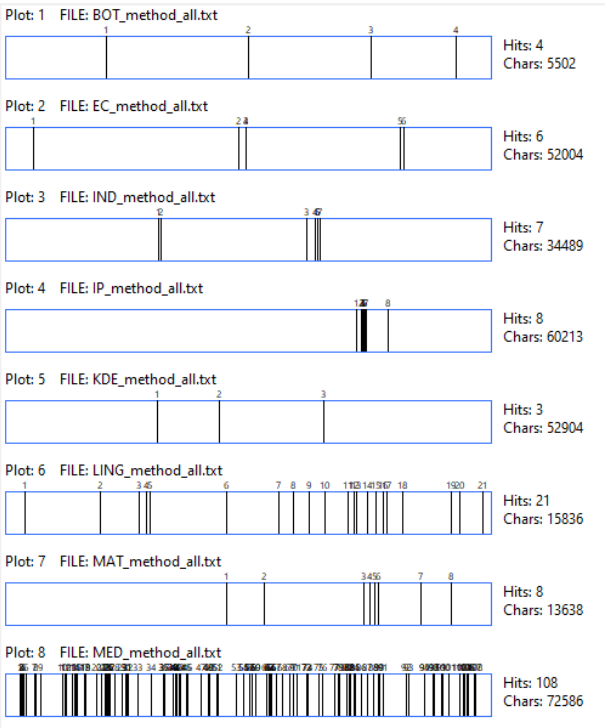


FIGURE 6.12: Concordance plot for regex search for regular verbs in passive voice in METHOD MOVE
[a] Plotted by AntConc 3.5.9

The rank frequency of the top three grammatical tenses differed by move and by discipline as shown in Tables 6.13 and 6.14. Disciplinary variation was also discovered for passive voice, which confirms that grammatical choice is affected by discipline.

Although there were some similarities between some disciplines and some moves, the lexical realization for all moves within each discipline is dissimilar. Thus, the null hypothesis $\mathbf{H}_0$ is rejected and the alternative hypothesis $\mathbf{H}_A$ is accepted.

6.4 Sub-question 9: Extent of move-specific lexical realization

6.4.1 Preamble

Sub-question 9 is:

> "To what extent does the lexical realization differ between moves?"

This research question builds on Sub-question 7 (Section 6.2). The question is exploratory and so there is no null hypothesis to be tested. The extent to which lexical realization differs is assessed by analyzing the relationship between rhetorical moves and keyness and tenses using cluster analysis.

6.4.2 Moves in keyness feature space

Keyness values were identified with the Keyness-calculator (The script is available in Appendix A.9) using the simple math formula (Kilgarriff, 2009) for the four primary

parts of speech namely adjective, adverb, noun and verb. To compare the relative usage of words among disciplines, ideally words which are used in all the disciplines are the best choice. This is because words that only occur in one or two disciplines cannot show finely grained distinctions between disciplines. For example if a word is used in the BOT sub-corpus but no other sub-corpus, the resultant data is binary, showing only two states. However, if the word occurs in multiple corpora then the count of the occurrence can be used. This enables comparison to made at numerous levels. Thus, to investigate the extent to which lexical realization is discipline specific, we do not want to focus on words that are highly technical and likely to be specific to only one research discipline, nor do we want to focus on words that are common to all genres, such as the highly frequent grammatical or functional words. Table 6.15 shows the relative frequencies of key words by part of speech. The most prevalent category is nouns with a total of 2,330 words. The number of adjectives stood at 818 while the number of verbs was 623. Nouns and adjectives tend to be technical and/or polysemous. The likelihood that the majority of the nouns, adjectives and probably verbs are discipline specific was felt to be high.

TABLE 6.15: Key words by parts of speech

Part of speech	Number of key words
Adjective	818
Adverb	196
Noun	2330
Verb	623

[c] The data was extracted from the results of a tailor-made script available in the Appendix A.9

Adverbs, however, were judged to be the part of speech with the highest likelihood of occurrence of the same tokens in multiple disciplines. For example, adverbs such as those of intensity (e.g. *very, too* and *extremely*), frequency (e.g. *usually, often* and *rarely*) and possibility or opinion (e.g. *probably, possibly* and *perhaps*) are likely to occur in many if not all of the disciplines.

A dispersion plot for the regex search for words ending in the suffix -ly resulted in 2,651 hits. Some of the hits were not adverbs, such as *supply*. However, as an indicator of the likely frequency of adverbs, manual inspection of the results showed that most concordance hits were, in fact, adverbs. The mean number of adverbs per sub-corpus is 54.3 with 48 out of 49 sub-corpora containing at least one hit. The smaller sub-corpora contained fewer hits. Ten sub-corpora contained ten or fewer hits while one sub-corpus (IP-discussion) contained no hits whatsoever. Figure 6.13 shows a typical extract of the KWIC view in AntConc 3.5.9 in which almost every instance is an adverb with the majority of adverbs being applicable to multiple disciplines. Some adverbs, such as *ectopically* [BOT], *galvanically* [IND], *magnetically* [IND], and *clinically* [MED], are likely to only occur in single disciplines given their specificity. Commonly-occurring adverbs, such as *finally, only, recently* and *significantly*, appear in multiple disciplines. Ideally, each sub-corpus of sentences (classified by discipline and rhetorical move) contains instances of the same tokens. There is no sub-corpus

FIGURE 6.13: Extract of KWIC[a] regex search to find words ending in -ly [b] in the whole corpus

[a] Key Word In Context concordance lines in AntConc 3.5.9

[b] Words coloured blue end in -ly and are likely to be adverbs

for MED-introduction as this is an empty set. Forty-nine sub-corpora were created: one for each discipline-move combination. Having established the suitability of the selection of the key words that are adverbs, a cluster analysis was conducted using the tailor-made script available in Appendix A.8.

Figure 6.14 shows the results of the *k*-means cluster analysis of rhetorical moves in keyness feature space using a fixed seed. Five distinct clusters were formed. The RESULT MOVE is located close to the origin (0, 0) while the remaining four moves are distributed in the four cardinal directions. This shows a lack of similarity in keyness among the adverbs in each of these rhetorical moves. This shows that adverb use is disciplinary specific. The most similar clusters judging by Euclidean distance were the RESULT MOVE and the PURPOSE MOVE, which were automatically grouped into the same category (and both colorized red in the Figure 6.14).

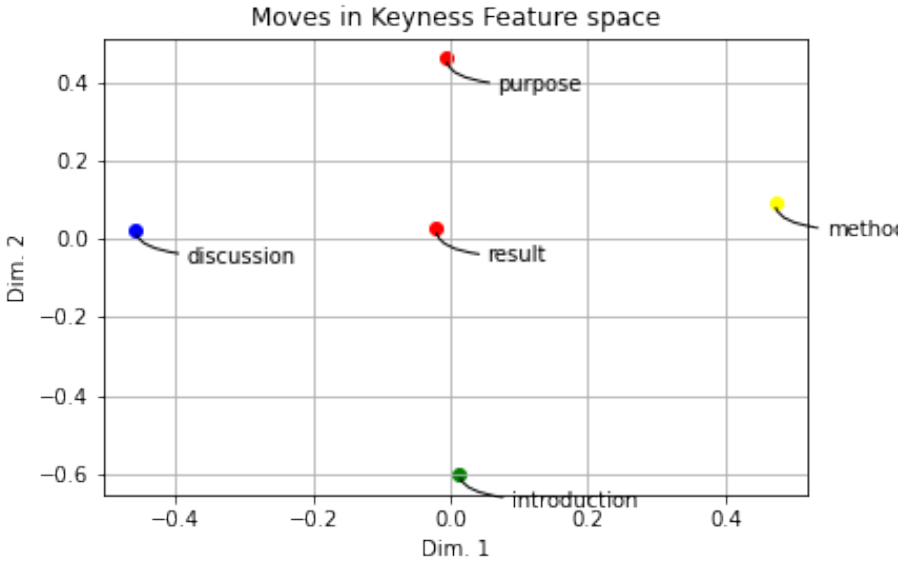


FIGURE 6.14: Plot of rhetorical moves in keyness feature space using fixed seed

To confirm that veracity of this result, another cluster analysis was carried out with the same variables but using a random seed. Figure 6.15 shows the results of the cluster analysis.

The distance between each of the clusters is approximately the same although the absolute location in the feature space differs. The overall pattern, however, remains the same with the RESULT MOVE located at the origin and the other moves at the cardinal points albeit rotated by approximately 180 degrees around the origin. The result shows that regardless of starting seed, the same clusters are formed.

6.4.3 Moves in tense feature space

K-means cluster analysis was carried out to see the relationship between rhetorical moves and grammatical tenses (See Appendix A.8 for the script). Figure 6.16 shows

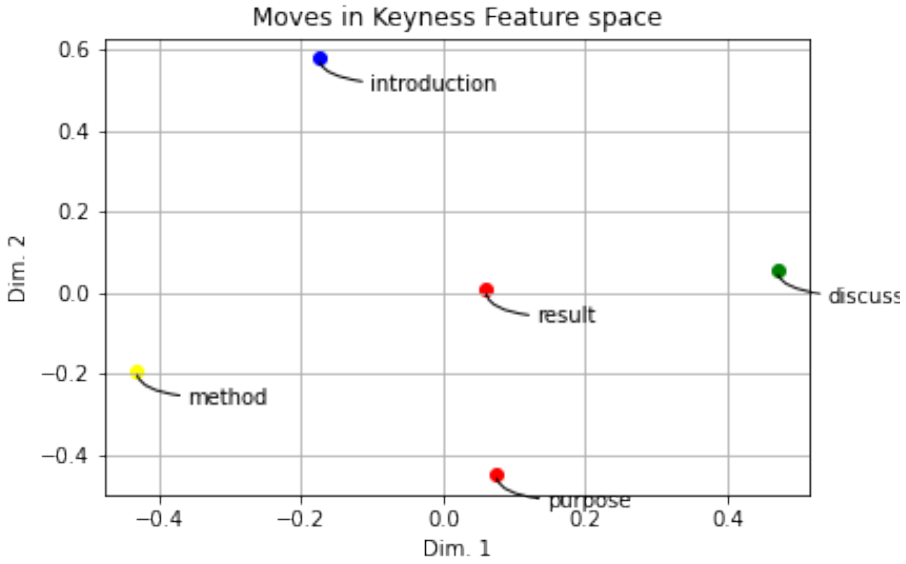


FIGURE 6.15: Plot of rhetorical moves in keyness feature space using random seed

the *k*-means cluster analysis plot of rhetorical moves in tense feature space using a fixed seed. In this plot, the rhetorical moves are dispersed, but not as widely as when plotted in keyness feature space. One quadrant of the tense feature space is unused. Although the METHOD MOVE and DISCUSSION MOVE are separated, they were automatically grouped into the same cluster as evidenced by the colouring. Perhaps, the algorithm identified a pattern of usage of *future perfect* and *present perfect* that were similar in these moves.

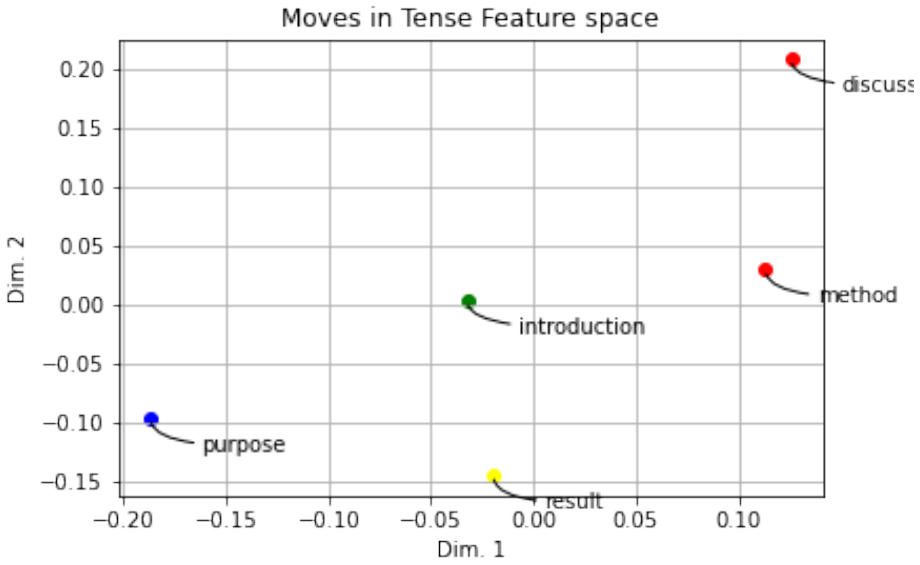


FIGURE 6.16: Plot of rhetorical moves in tense feature space using fixed seed

6.4.4 Moves in keyness and tense feature space

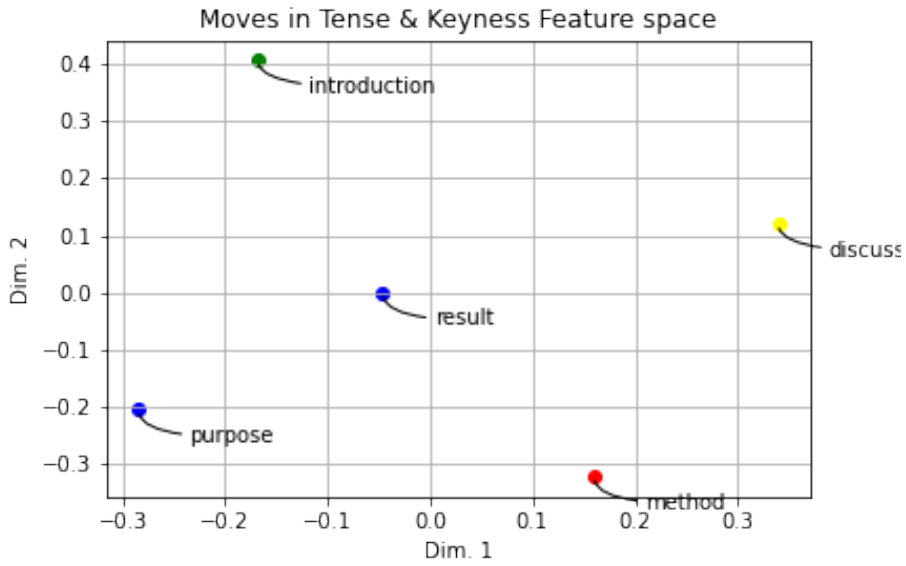


FIGURE 6.17: Plot of rhetorical moves in tense and keyness feature space using fixed seed

Rather than relying solely on either keyness or tense, a *k*-means cluster analysis of rhetorical moves in tense and keyness feature space was carried out with the same script and a fixed seed. Figure 6.17 shows the *k*-means cluster analysis plot of rhetorical moves in tense and keyness feature space. The overall distribution of clusters is similar to the initial cluster plot for rhetorical moves in keyness feature space with the RESULT MOVE remaining close to the origin, but the clusters for the four other moves appear to have rotated around 20 degrees anticlockwise and the PURPOSE MOVE and METHOD MOVE appear to have swapped positions. What is notable, however, is that in this plot the RESULT MOVE and PURPOSE MOVE are now grouped into the same cluster, colorized blue.

During the annotation phase, differentiating between the RESULT MOVE and PURPOSE MOVE was problematic. A case in point is the common usage of the word *propose* to announce results. However, *propose* was also used to introduce the general aim or goal with specific results being stated explicitly later in a RESULT MOVE. In all the published research abstracts analyzed intensively, proposed algorithms and systems were shown to improve on existing algorithms and systems by some metric. The lack of published negative results may, therefore, lead readers to assume that any proposal with no specific announcement of results is not just a proposal, but that it will be followed by the announcement of positive results in the accompanying research article.

Discussions with specialist informants confirmed this. One specialist informant, who served as an editor for one of the top-tier journals in this study, noted that negative results would not be published in the journal that he represented and so the

proposal statement could be considered as a framing device leading readers to infer that positive results follow. If the proposal statement was not followed by specific results, the proposal would be considered as a RESULT MOVE. Discriminating the RESULT MOVE from a PURPOSE MOVE was challenging in the annotation process, and so this clustering result can be understood from that perspective.

From the perspective of a grammar-orientated teacher of writing though, it raises questions since different tenses are usually advocated for these moves with *present simple* being advocated for the PURPOSE MOVE and *past simple* for the RESULT MOVE. However, as was shown by the rank frequency of tenses by move, *past simple* ranked the most frequent tense in the RESULT MOVE only in the MED corpus.

6.4.5 Sub-question 9 conclusion

In sum, when rhetorical moves were plotted in the same vector spaces, the moves were distinct and separate regardless of features selected. This shows that lexical realization judged by keyness and grammatical tense differs between rhetorical moves. The difference between the lexical realization in each move were so great that each of the five moves plotted were spread apart over the keyness feature space in the plot of the cluster analysis. Although *present simple* form is the most frequent grammatical tense, there are variations in the rank frequency and distribution frequency of grammatical tenses across rhetorical moves.

Lexical realization differed in both keyness and grammatical tense, but keyness appeared to show more distinct differences based on the plots of the results of the cluster analysis. This analysis used the keyness of adverbs but no doubt should nouns or verbs have been selected, fewer similarities would have been uncovered due to the highly technical terminology in a number of disciplines. The results of the cluster analysis confirm the results described in Sub-question 6.3. Overall, given the high degree of distinctive correlations between lexical realization and rhetorical moves, it can be concluded that lexical realization is greatly affected by or greatly affects rhetorical moves.

6.5 Sub-question 10: Extent of discipline-specific lexical realization

6.5.1 Preamble

Sub-question 10 is:

> "To what extent does the lexical realization differ between disciplines?"

This research question builds on Sub-question 7 (Section 6.2). The question is exploratory and so there is no null hypothesis to be tested. In the same way as Sub-question 9 (Section 6.4), the extent to which lexical realization differs is ascertained by

using cluster analysis to investigate the relationship between disciplines and lexical realization, using the proxies of keyness and grammatical tenses.

6.5.2 Disciplines in keyness feature space

We can discover any similarities among the lexical realization in disciplines by exploring how the disciplines are related to the feature spaces of keyness and tense individually and jointly.

Figure 6.18 shows two *k*-means cluster analysis plots of disciplines in tense and keyness feature space. Figure 6.18a shows the *k*-means cluster analysis using a fixed seed while Figure 6.18b shows the results using a random seed.

The automated colorization resulted in the same clusters although their absolute location in the feature space differed. An unexpected result was the inclusion of linguistics in the same cluster as IND, IP, IT, KDE and WC. In both plots, the MED and BOT abstracts formed clusters of one discipline each. The Euclidean distance between BOT and MED was the greatest. MAT and EC were grouped together and so the usage of relative frequency of adverbs in both disciplines must share similarities.

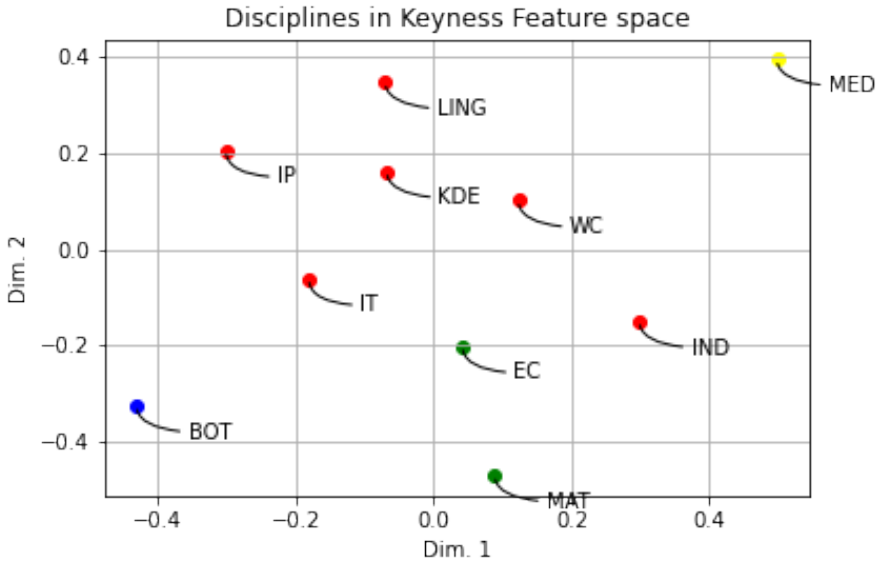
6.5.3 Disciplines in tense feature space

Figure 6.19 shows the *k*-means cluster analysis plot of disciplines in tense feature space using fixed and random seeds. Figure 6.19a shows the cluster plot for the fixed seed while Figure 6.19b shows the plot for the initial random seed.

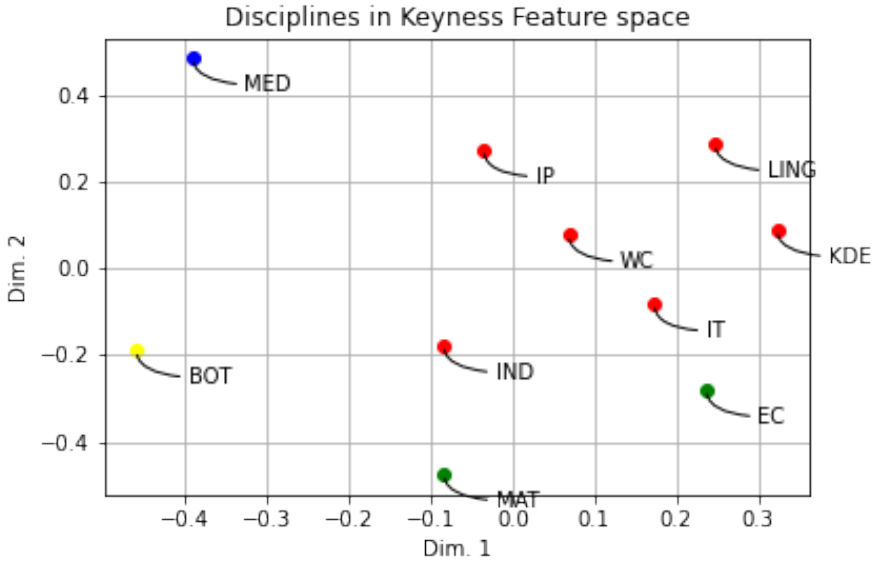
When plotted in the tense feature space, the disciplines appear much more similiar. In fact, the points for some disciplines almost overlap. Notably, as with the plot of disciplines in keyness feature space, MED stands apart from all other disciplines. In the tense feature space, the disciplines of BOT and LING are grouped together. This could be due to similar distribution patterns of grammatical tenses.

The red cluster of six disciplines comprises all the disciplines that may be grouped under the umbrella term information science. This grouping shows that the underlying patterns of grammatical tense used are similar across these disciplines. Both BOT and LING are categorized into a single cluster in the tense feature space. This grouping may be explained by the fact that both the LING and BOT corpora have similar distributions of *present simple*, *past simple* and *future simple* forms.

Overall, there appear to be more similarities between disciplines in terms of grammatical tense than keyness, judging by the much closer clustering of disciplines. With the benefit of hindsight, this would be expected since the set of grammatical tenses contains only twelve elements, and follows a Zipfian distribution. With only three grammatical tenses used in 90% of the corpus, it would not be unusual to find that most disciplines follow a similar overall distribution. However, the MED corpus in particular stands out as not doing so. The medical abstract corpus was the only one with four times more instances of *past simple* than *present simple*.

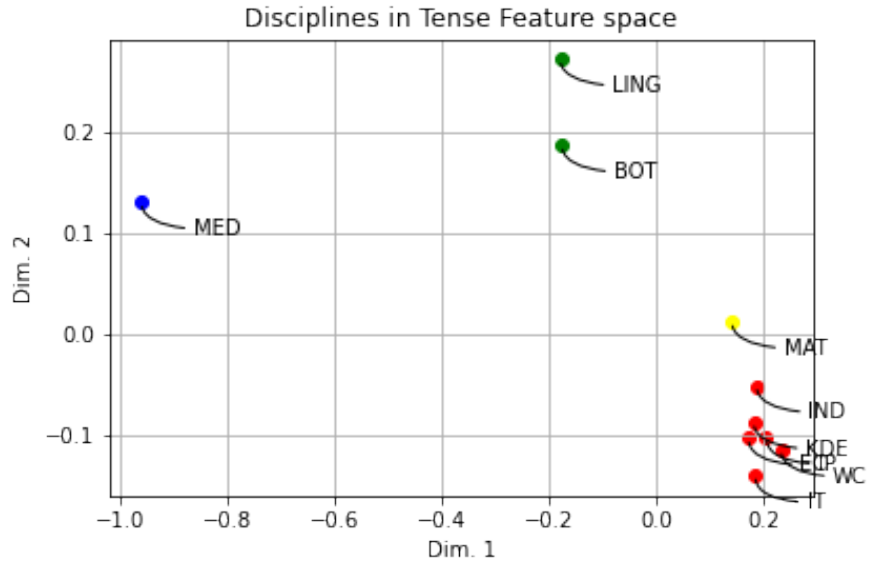


(A) Fixed seed

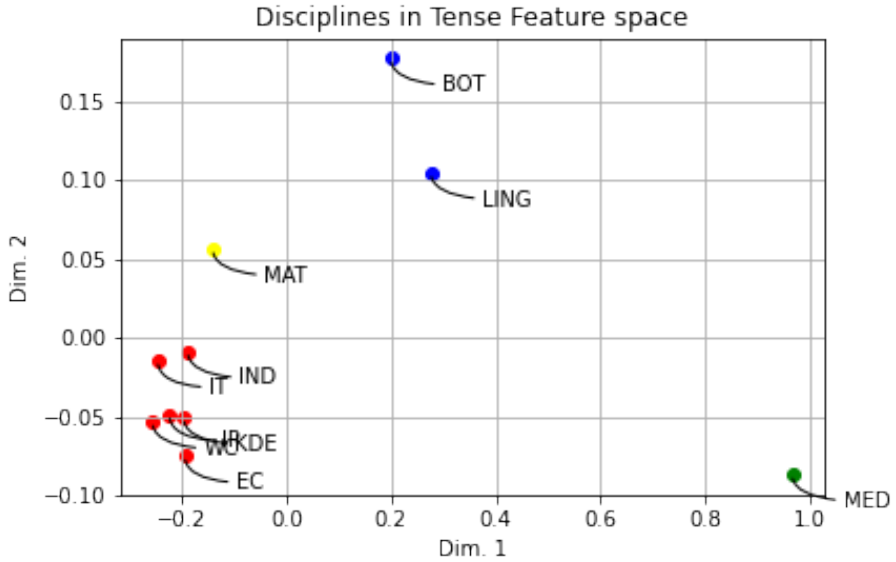


(B) Random seed

FIGURE 6.18: Plot of disciplines in keyness feature space



(A) Fixed seed

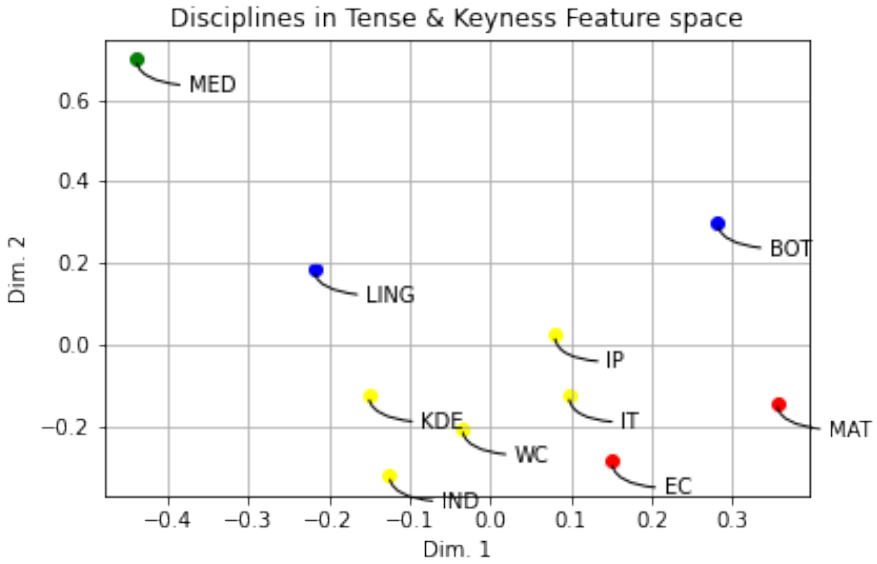


(B) Random seed

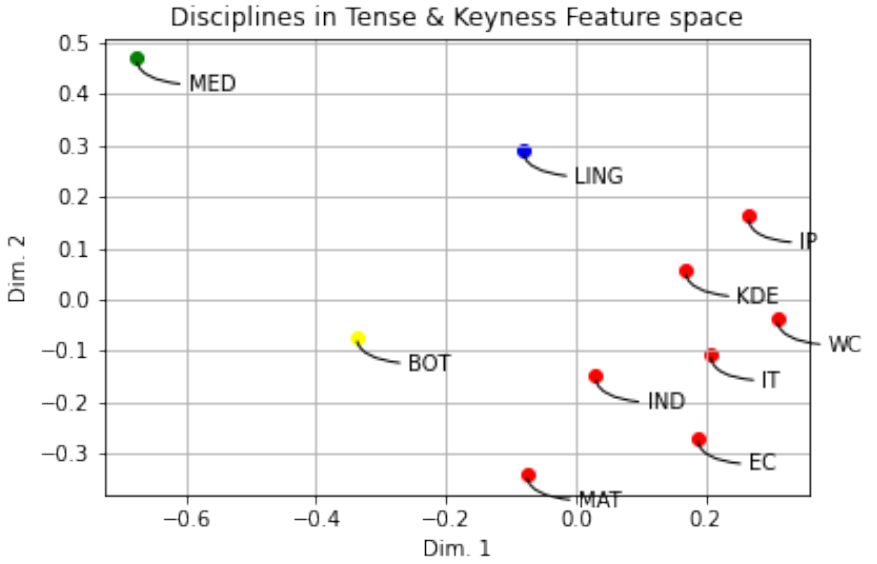
FIGURE 6.19: Plot of disciplines in tense feature space

6.5.4 Disciplines in keyness and tense feature space

Figure 6.20 shows the *k*-means cluster analysis plot of disciplines in tense and keyness feature space. Figure 6.20a shows the cluster plot for fixed seed while Figure 6.20b shows the plot using a random-seed.



(A) Fixed seed



(B) Random seed

FIGURE 6.20: Plot of disciplines in tense and keyness feature space

In these cluster plots, the disciplines are spread across the feature space and automatically grouped into clusters. The single discipline cluster of MED is the most distant from the other disciplines. The main group of applied engineering disciplines are close together, but were automatically categorized differently depending on the starting seed. The random seed cluster plot in Figure 6.20b is probably the easiest

to follow. In this plot, both BOT and LING are grouped into cluster comprising one discipline each with the remaining seven disciplines forming one large cluster. When compared with Figure 6.20a, the disciplines of EC and MAT are separated out into their own individual clusters.

6.5.5 Sub-question 10 conclusion

If the lexical realization is mainly determined by move, the lexical realization should be similar regardless of discipline. However, Subsections 6.4.2 and 6.4.3 have shown that there is considerable variation in lexical realization across disciplines even for the same rhetorical move.

The results of the *k*-means cluster analysis show that when plotting moves in keyness feature space, the moves are spread apart rather placed equidistantly. The PURPOSE MOVE and RESULT MOVE shared the most similarity. Given the difficulty of demarcating these categories during the annotation stage, this result makes sense. Proposal statements (e.g. This paper *proposes* . . .) at times may appear to be announcing results and therefore be mistaken as RESULT MOVE. The similarity revealed in the cluster analysis therefore reflected the reader interpretation of the move in context. What is slightly surprising is the lack of similarity between the INTRODUCTION MOVE and the PURPOSE MOVE when viewed in keyness feature space. This may be due to the wide range of lexis that may be included in the much broader category of INTRODUCTION MOVE.

In short, there seems to be more similarity in lexical realization across rhetorical moves in the same discipline than across disciplines in the same move. Put simply, discipline-specific lexical choice appears to trump move-specific lexical choice.

6.6 Implications and applications

6.6.1 Preamble

This section summarizes and extends the theoretical implications that have been discussed in this chapter so far. The practical applications are also directly linked to the implications.

6.6.2 Generic integrity

In terms of rhetorical move sequencing, the results described in Chapter 5 showed that IMRD and IPMRD linear sequences were unrepresentative of move permutations occurring in this corpus. This knowledge will surely help novice writers more quickly realize the lack of necessity to conform to any prescribed linear sequence. In terms of lexical realization, perhaps, the biggest take-away is that lexical variation appears to be more determined by discipline than by move.

The fact that discipline-specific lexical choice appears to surpass move-specific lexical choice is good news for scientists with a reasonable grasp of English. Scientists

are more likely to be familiar with the technical terminology in their discipline than with any terminology related to rhetorical moves.

In some disciplines, some rhetorical moves are so conventionalized that boilerplate text is used. This type of text is both discipline-specific and move-specific. This is the case in the METHOD MOVE for materials science in which different variables (e.g. materials, quantity, temperature, time, etc.) are plugged into template-like sentences; or in information theory when describing mathematical proofs.

This is, however, not good news for teachers of scientific writing with very limited knowledge of the disciplines of their students. If move-specific terms were more similar, it would be easier to generalize the type of lexical and grammatical choices used based purely on the rhetorical move without considering disciplinary variation in detail. This is not the case. Teachers without disciplinary knowledge therefore either need to (1.) check the suitability of their advice themselves – ideally before providing the advice or (2.) ask students to check the suitability of their advice against published abstracts.

A central problem for teachers with little disciplinary knowledge is the difficulty of understanding abstracts without any knowledge of technical terminology. It has been argued that disciplinary knowledge is not needed to teach scientific writers [2]. However, this is a disservice to the students. Teachers need not be disciplinary experts but they need to be able to understand the meaning of the texts enough so as to give the best quality advice. Being able to read an abstract and know that, for example, concept A affects element B and causes effect C is not understanding the abstract. It might be that in the real-world concept A does not affect element B, but is affected by element B. Teachers of writing without disciplinary knowledge usually note that it is the responsibility of the research supervisor to identify such mistakes in meaning; yet this mistake may simply be an error in selecting active voice rather than passive voice. The research supervisor, however, could easily argue that the error is purely a language mistake and not a scientific error, and so should fall within the duties of teachers of research writing.

This disciplinary knowledge may be garnered through corpus enquiry, reading relevant discipline-specific textbooks or through extensive exposure to research articles.

6.6.3 Collocations

Numerous researchers (Barnbrook, O. Mason, and Krishnamurthy, 2013; Haswell, 1991; Hyland, 2008; Paice, 1980; Sinclair, 1991) have argued that language is constructed by sequencing chunks (blocks or strings) of words. Co-occurrence patterns involve idioms and fixed expression to looser coupling of co-occurrence of words (Dalpanagioti, 2018) , which depending on the words selected may result in *marked* or

[2] Anthony, L. (2011, October). Why ESP practitioners do NOT need to be subject specialists. Keynote speech presented at the 2011 International Conference and Workshop on English for Specific Purposes (ICESP 2011), HungKuang University, Taichung, Taiwan.

unmarked forms. The collocations used in these chunks are affected by discipline and register.

Disciplinary collocations may be highly specialized and relate to only a single discipline and even a single specialism within a discipline. For example, even though the adjective *big* is considered as less formal than its counterpart *large*, in information science the accepted term to describe large volumes of data is *big data* and not *large data*. This is the disciplinary impact of lexical choice. When helping postgraduate researchers to write up their research, it is likely that they will have mastered the technical terminology. This is for two main reasons. First, the same technical terms (albeit pronounced differently) are often used across languages even those with different orthographies. Second, most students have already studied their first degree in the same disciplinary field and will have had extensive exposure to the technical terminology. To provide a linguistic example, all linguistic researchers would be expected to understand the multi-word expression *part of speech* but only those in systemic functional linguistics would be expected to understand *textual metafunction*[3].

Scientific research abstracts are written in a formal register, eschewing any hints of informality (e.g. undefined abbreviations, contractions, non-standard forms, etc.). The research did not directly focus on register, but from a practical perspective, it is a relatively easy concept to convey to novice writers and is also easy for writers to systematically incorporate.

Disciplinary variation could be discovered by exploring the key words. Identifying trigrams and 4-grams that frequently occur within particular rhetorical moves would certainly help raise learner awareness of collocation. As was shown in Figure 6.10, the term *randomised controlled trials* occurred in medical abstracts but did not occur in any other disciplines.

The lexical patterns within each move within each discipline can be used for pedagogic purposes.These patterns can be combined to create skeleton sentences from which writers can select grammatically accurate choices appropriate to their discipline. Substitution tables can be used to provide model sentences for writers to select from. This provides writers with a variety of choices that conform to a particular pattern. Substitution tables provide scaffolding for novice writers and are commonly used for learners with English as an Additional Language.

Examples 43, 44 and 45 show how novice writers can be provided with model sentences based on the lexical realizations discovered in this study, yet be given some flexibility to tailor the sentence to their research field and discipline. These corpus-based examples show how the OVERVIEW SUB-MOVE can be realized in the INTRODUCTION MOVE.

(43) The (X / Y) of (the / this) paper is (organized / structured) as follows:

(44) Section (One / 1) describes the (background / motivation / problem) for this study.

[3]Textual metafunction focuses on the grammatical systems responsible for managing the flow of discourse

(45) This paper (ends / finishes) with a (brief / concise) (conclusion / summary).

6.6.4 Colligations

Colligation (Hoey, 2005; Gledhill, 2009) between grammatical tense and rhetorical moves are evident in the corpus. The overall ranking of the grammatical tenses follows a similar distribution pattern as discovered by Biber, Johansson, et al. (1999) in spoken English. However, patterns of usage of grammatical tense were discerned when analyzing the corpus and comparing the discipline-move sub-corpora. This showed that colligation of grammatical tense was affected by both variables.

Grammatical tenses show colligation with rhetorical moves and discipline. In some disciplines, only one grammatical tense for a particular move was discovered. Materials science is a case in point. *Present simple tense* was the only tense discovered in the INTRODUCTION MOVE and PURPOSE MOVE. In other disciplines although three or more grammatical tenses were found in a rhetorical move, one grammatical tense far outweighed the others. For example, in the METHOD MOVE for wireless communication, there were 209 instances of *present simple tense* compared to 9 instances of other tenses. However, in the botany corpus, the usage of grammatical tenses was less clear cut in the RESULT MOVE with 210 instances of *present simple tense* to 202 instances of *past simple tense*.

Although writers need to consider the grammatical meaning of the tense, it appears there is little need to understand the fine nuances of multiple tenses.

In fact, apart from the METHOD MOVE in BOT, LING and MED and the RESULT MOVE in LING and MED, the generic advice to a novice writer that is most likely to result in an abstract that is conforms the generic expectations in a discipline is to use *present simple*.

Given that slightly under 70% of all the grammatical tenses discovered in the corpus were *present simple*, recommending this tense as a default option is a reasonable pragmatic choice. In fact, bar the five move-discipline sub-corpora mentioned above, *present simple* was the top ranked grammatical tense for each rhetorical move in each discipline.

This single default option could be refined by adding a second finer-grained level of advice which allows three simple tenses, novice writers should be able to draft an acceptable abstract. This results in an easy-to-follow decision tree for tense selection, such as shown in Figure 6.21.

This decision tree is obviously an oversimplification; nevertheless, it should help novice writers get their first draft of an abstract without struggling over which tenses to use. The choice of tenses needs to be considered again more carefully once the initial draft is complete. However, in terms of coverage, based on the corpus results, statistically *present simple* tense covers approximately 70% of all cases, while the three simple tenses cover over 90% of all cases. To cover the remaining cases the other nine grammatical tenses need to be considered.

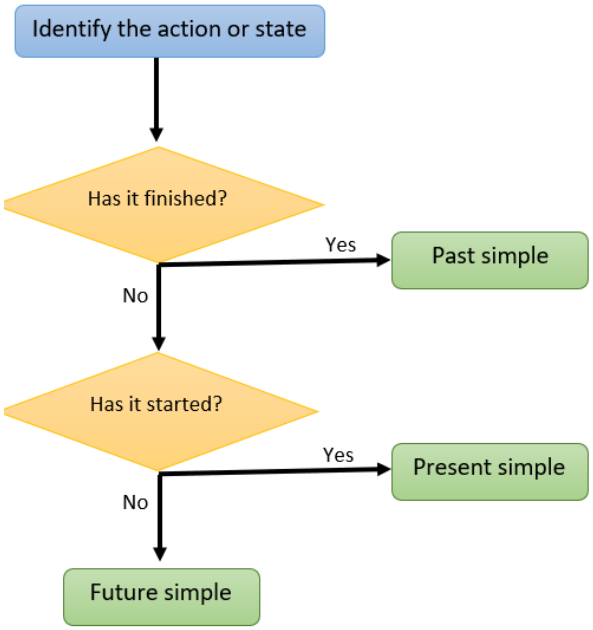


FIGURE 6.21: Decision tree for three simple tenses

Probably, the best way to help novice writers is to understand move-specific tense colligation is to show examples of tenses used in context.

Figure 6.22 shows an easy-to-follow medical abstract that was published in the British Medical Journal. The finite verbs have been automatically detected by a prototype tense identification script (Blake, 2020a)[4]. This script executes a visualization function that colorizes each tense according to the key given.

There are, however, some false positive results caused by erroneous tagging errors. Thirteen verbs were identified and accurately labelled. Two verbs were labelled incorrectly and one word, *ratio*, was labelled incorrectly. The word *ratio* is polysemous and may be a verb in some contexts.

Despite the tagging errors, this tool can be used to discuss the use of tenses in context. According to the noticing hypothesis (Schmidt, 2012), learners do not learn what they do not notice; and so any activity that raising awareness of colligation may act as a vehicle for learning.

6.7 Chapter Summary

The lexical realization within rhetorical moves differed in terms of both lexis and grammar when measured using key words as a proxy for lexis and grammatical tenses as a proxy for grammar. Lexical realization differed between the same moves in different disciplines, and between different moves in the same discipline.

[4]This tense identification and visualization tool is available online at `https://tense-sense-identifier.herokuapp.com/`

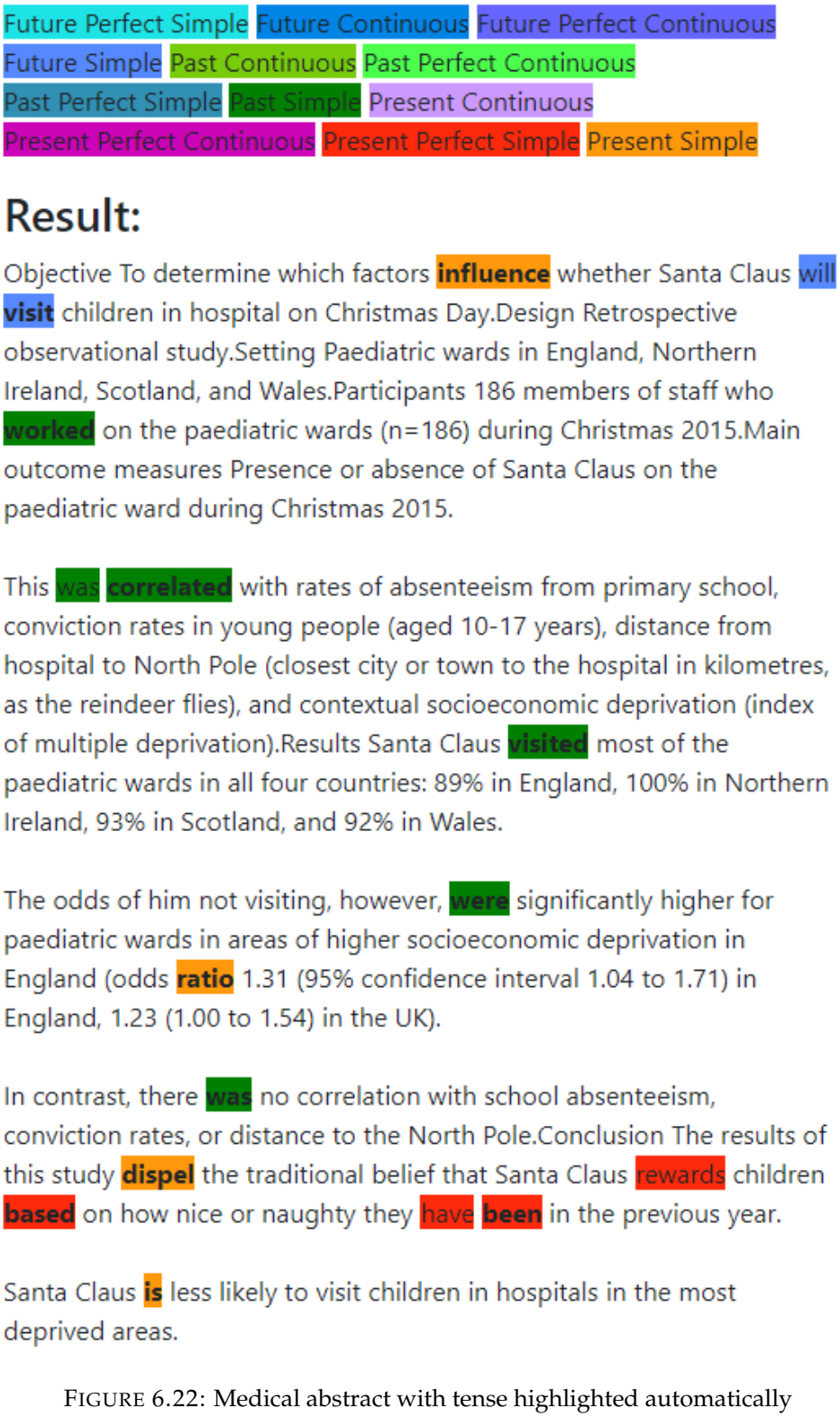


FIGURE 6.22: Medical abstract with tense highlighted automatically using prototype script

When investigating the dispersion of the top ten key words for the corpus, it was found that the key words were distributed unevenly. In fact, the corpus-wide key words did not occur in some moves for some disciplines.

The investigation of grammatical tenses showed that *simple* tenses account for almost all (96%) of the grammatical tenses in this corpus. Present simple was ranked the most frequent tense is most of the discipline-move sub-corpora. Although the wide-spread usage of simple tenses permeated the corpus, the dispersion of tenses within moves differed between disciplines. Disciplinary variation was also discovered for passive voice, which confirms that grammatical choice is affected by discipline.

If the lexical realization is mainly determined by move, the lexical realization should be similar for the same move regardless of discipline. However, this was not the case, and so there appears to be more similarity in lexical realization by discipline than by move. This effect was most notable when comparing and contrasting the key words. Many of the top five ranked key words within a discipline occurred in multiple moves within the same discipline. The same set of key words occurred far less frequently in the same move in different disciplines. The dispersion of the top ten key words in the same moves also differed between disciplines. The plot of the *k*-means cluster analysis for rhetorical moves in three different vector spaces confirmed that rhetorical moves were distinct and separate. The differences in Euclidean distance between move clusters in keyness feature space exceeded that of tense feature space.

Drawing on ideas from the Sydney school of SFL, teachers encourage novice writers to deconstruct, reconstruct and create scientific research abstracts. Through the interaction between teacher and student, and student to student; novice writers can increase their awareness of the generic conventions, and then apply this newly-gained knowledge. Online visualization software for rhetorical moves and grammatical tenses could be utilized. Novice writers can access an abstract in the *Move Visualizer* and show or hide the moves on demand. An online tense identifier that colorizes verb forms in context by their grammatical tense could be a useful tool to raise awareness of tense-move colligation. Noticing is a necessary precursor to understanding and so activities such as identifying tenses and rhetorical moves in context enable learners to climb the learning curve.

Overall, given the high degree of distinctive correlations between lexical realization and rhetorical moves, it can be concluded that lexical realization is greatly affected by or greatly affects rhetorical moves. There seems to be more similarity in lexical realization across rhetorical moves in the same discipline than across disciplines in the same move. In short, discipline-specific lexical choice trumps move-specific lexical choice.

Chapter 7

Conclusion

To do successful research, you don't need to know everything, you just need to know one thing that isn't known.
— Arthur L. Schawlow[1]

7.1 Chapter preview

Move analysis of conclusions in research articles in applied linguistics has shown that typical rhetorical moves include summarizing the study, evaluating the study and making deductions from the research (R. Yang and Allison, 2003). All three rhetorical moves can also be found in this chapter.

Section 7.2 provides a concise summary of the research process from the initial literature review through to each of the three phrases in the methodology.

The corpus was investigated, language features were identified, extracted and analyzed. The result, a description of the corpus of scientific research abstracts, is given in Section 7.3. The description starts by outlining the corpus dimensions. The rhetorical organization in terms of moves and move sequencing is described. The lexical realization within the rhetorical moves is then described. Notable features relating to rhetorical moves, lexical realization and disciplinary variation are discussed.

The specific language features identified during this study are focussed on in more depth in Section 7.4. This section begins by itemizing the propositions that were formulated related to rhetorical organization and lexical realization within moves. These propositions are supported by evidence from the corpus. Arguments using the propositions are then used to generate conclusions, which are referred to as *claims*. The evidence, premises, assumptions and reasoning used to generate conclusions and any intermediate conclusions are clearly stated where appropriate.

Section 7.5 provides a concise summary of the answers to the research questions regarding rhetorical organization while Section 7.6 deals with the research questions regarding lexical realization within rhetorical moves.

Research without any real-world application is purely theoretical. This research, however, aims to impact the teaching of research writing. Section 7.7 shows how the

[1] As quoted in Steven Chu and Charles H. Townes, 'Arthur Schawlow', Biographical Memoirs of the National Academy of Sciences (2003), Vol. 83, 7.

theoretical results may be applied to language learning, particularly for novice writers with English as an additional language who are either studying research writing or need to draft scientific research abstracts. Ways in which teachers of scientific writing can also harness the results are given. Some of the potential applications are described and specific examples provided.

Limitations are described in Section 7.8. The three primary limitations are the size of the corpus, the veracity of the annotations and the accuracy of the tailor-made software program to identify grammatical tenses.

Five suggestions for future work are provided in Section 7.9. Two of the suggestions focus on increasing the accuracy of automatic language feature identification while the other three focus on corpus approaches.

7.2 Summary of research process

This investigation on rhetorical organization of scientific research abstracts has built on the work of many researchers, but three works that influenced this research the greatest are Bhatia (1993), Hyland (2004), and Swales (1990).

The review of the literature revealed that most research on rhetorical organization focussed on the disciplines of linguistics and medicine. Little research had been conducted on disciplines in applied sciences. Researchers tended to report the presence or absence of rhetorical moves rather than describe patterns of rhetorical moves. Given the paucity of corpora annotated for rhetorical moves, little was known about the lexical realization within moves. Although many researchers had extracted lists of keywords and multi-word expressions for a limited number of corpora, no studies were discovered that investigated patterns of grammatical tense usage across moves in different disciplines. Two main research questions were formulated to fill the research gap revealed in the literature review. The questions were:

1. What is the rhetorical organization of abstracts of research articles published in a broad range of top-tier scientific journals?

2. What are the lexical features of prototypical moves in abstracts of research articles in the selected scientific disciplines?

To investigate these research questions, a corpus-based approach using a specialist corpus was selected. A corpus approach was selected because of the benefits of using a statistical analysis of language features occurring in a dataset. This approach is evidence-based, replicable and harnesses extrospection. To identify hidden patterns, it was necessary to annotate the corpus by hand for rhetorical moves. This added value to the corpus since analysis of language features could be investigated between both discipline and rhetorical move.

A three-phase method was adopted comprising corpus, annotation and analysis phases. In the corpus phase, a balanced corpus of scientific research abstracts was

compiled including under-represented hard-to-read disciplines. During the annotation phase, a detailed annotation protocol was developed. The principal investigator annotated the whole corpus. The veracity of the annotation of rhetorical moves was confirmed through double annotation and the use of specialist informants. In the analysis phase, numerous programs were developed to extract and analyze the data. Lexical realization was investigated using keyness as a proxy of lexis and grammatical tense as a proxy for grammar. The similarity and differences in lexical realization were analyzed using *k*-means hierarchical clustering using tense and keyness separately and together.

The results of the corpus-based study are summarized in Section 7.3. The corpus description is intended to be simply that. The theoretical claims that stem from the corpus results and are mentioned specifically or alluded to in the corpus description are itemized in the two subsequent sections. Section 7.4 describes the theoretical claims that can be made based on corpus evidence related to rhetorical organization and lexical realization within rhetorical moves. Sections 7.5 and 7.6 provide the answers to the research questions that arose from the literature review.

7.3 Corpus description

An isotextual corpus of 1000 research abstracts comprising ten different scientific disciplines [BOT, EC, IND, IP, IT, KDE, LING, MAT, MED and WC]. The disciplines were selected from each branch of the taxonomy proposed by Becher and Trowler (2001). One important feature of this corpus is the selection of under-represented scientific disciplines, such as wireless communication and information theory. The total number of word tokens was slightly under 170,000 (n = 169,499) of which there were slightly over 12,000 word types (n = 12,248). In total there were 7200 sentences. A sentence was judged as a string that begins with a capital letter, comprises at least one unit of meaning and ends with a end stop, which in this corpus was always a full stop. The mean sentence length for the corpus was 23.5 with discipline means varying from approximately 22 to 26. When judged using the Flesch Kincaid grade level and Flesch Kincaid reading ease scores, readability for all disciplines was classified as very difficult for all disciplines apart from MED which was classified as difficult.

The corpus was examined in detail to understand the rhetorical organization and the lexical realization within the rhetorical moves.

Of the ten selected disciplines, two disciplines, MED and MAT, stood out as being different from the outset without the need for any statistical analysis. This was due to the difference in type of abstract. The other eight disciplines adopted traditional abstracts. MED abstracts were structured abstracts with pre-determined headings. Their mean length was approximately twice as long as the traditional abstracts. MAT adopted graphical abstracts comprising an image and text. Graphical abstracts were on average approximately half the length of traditional abstracts.

Four out of the five rhetorical moves (PURPOSE, METHOD, RESULT and DISCUSSION MOVE) occurred in all the disciplines. There were no instances of INTRODUCTION MOVE in medical abstracts. Rhetorical moves were not evenly distributed. Approximately 40% were RESULT MOVES, 30% METHOD MOVES, and 20% INTRODUCTION MOVES. PURPOSE MOVES and DISCUSSION MOVES stood at approximately 5% each. The distribution of the proportion of rhetorical moves varied by discipline. The most frequent initial rhetorical moves to begin abstracts varied by discipline. There was a pattern: the disciplines EC and KDE tended to commence with INTRODUCTION MOVES; the disciplines IP, IT and MED tended to use METHOD MOVES; and the remaining disciplines frequently started with RESULT MOVES.

The number of moves within abstracts ranged considerably with the shortest abstracts comprising only one move while the longest had nine moves. Four-move abstracts were the most frequent (n = 340) and accounted for slightly less than half of all the moves.

An investigation of the frequency of adjacent pairs of rhetorical moves threw light on the prevalence of non-linear sequences. Specifically, although the commonly expected order of the METHOD MOVE and RESULT MOVE is sequential, many instances were found of the moves occurring in reverse. In fact, in two disciplines, IP and KDE, the order METHOD-RESULT MOVE was only slightly more common than RESULT-METHOD MOVE.

Three dimensions were uncovered regarding sequences of moves, namely variation, linearity and cyclicity. First, there were more permutations of move sequences than initially envisaged and far more than have previously been described in the literature. Second is linearity. The literature tends to describe moves occurring in a linear sequence beginning with the INTRODUCTION MOVE and ending with the DISCUSSION MOVE. However, the results of this corpus study revealed that non-linear sequences were commonplace. Third, the phenomenon of move cycling was discovered. This involves the repetition of two or three sequential moves. The most common cyclical patterns were for the METHOD MOVE and RESULT MOVE.

The anticipated and well-documented rhetorical organisation structures of IMRD and IPMRD were found in the corpus. However, these two four-move and five-move permutations made up only a small fraction of the different permutations of rhetorical move sequences. Slightly under 200 different permutations of moves were discovered.

Although this number of permutations is far more than initially envisaged, it is far lower than the number of potential permutations based on calculations for three simple scenarios for potential permutations and investigating the actual number of permutations discovered in the corpus. In one scenario with 120 potential permutations, just three were realized.

It should be noted that in contrast to the linearity and cyclicity, variation cannot be judged from an individual abstract, but is evaluated based on the whole disciplinary corpus, which in this study comprises 100 texts.

Linearity is based on the expected sequence of rhetorical moves. This expected sequence according to the literature review for four moves was IMRD while for five moves the expected sequence was IPMRD. Discussions of non-linear abstracts with specialist informants revealed that authors tend to front the result move in some disciplines. It was also noted that this was more common with short research articles (e.g. letters or communications) in IND.

In disciplines like industrial engineering and wireless communication that create physical or abstract artefacts such as machines, circuits and algorithms; cyclicity was frequently found. For example, METHOD-RESULT MOVE could describe the development phase after which METHOD-RESULT MOVE could be used a second time to describe the evaluation phase. One effect of this cyclicity is the creation of a non-linear sequence between the moves describing the development and evaluation phases.

The *Borromean Rings* framework was created to enable the mapping of disciplines onto a framework based on the rhetorical organization of scientific research abstracts. This framework harnesses three interlocking circles, which creates eight discrete regions into which disciplines can be placed. Each region is allocated different values for the dimensions of linearity, cyclicity and variation.

7.4 Deductions and inferences

> *There are two modes of acquiring knowledge, namely by reasoning and experience. Reasoning draws a conclusion and makes us grant the conclusion, but does not make the conclusion certain, nor does it remove doubt so that the mind may rest on the intuition of truth, unless the mind discovers it by the path of experience.*
> — Roger Bacon, English philosopher

This section shows how data in the corpus is organised into information that can be transformed into knowledge. The first step is the creation of propositions founded on corpus data. These propositions may be combined and reasoning applied to reach conclusions. The conclusions, or claims, are limited to the corpus, and no attempt is made at generalizing beyond the corpus. However, the claims are phrased tentatively and so may be applicable to different corpora. The focus of the research is on helping non-writers and so these conclusions aim to describe prototypical moves and move sequences. Anomalies and outliers are, at times, ignored.

7.4.1 Propositions and arguments

The list of propositions that were formulated based on the corpus evidence is given in Table 7.1. References are provided to relevant sections of the thesis in which the evidence supporting each proposition can be confirmed. Some propositions are used as premises in logical arguments to which reasoning could be applied to generate conclusions or claims. All claims are not equal, though. Claims that are

based on deductive reasoning using premises which are true, and a logical argument that is valid may be termed sound conclusions. Claims that are based on inductive reasoning using true premises for which evidence is strong may be termed cogent conclusions. The key difference is that sound conclusions are certainly true while cogent conclusions may not be. However, should any premise be false, any argument be invalid or any evidence weak, conclusions are neither sound nor cogent.

TABLE 7.1: List of propositions on rhetorical organization and lexical realization in scientific research abstracts

Number	Proposition	Reference
1	Mean length of graphical abstracts is L^a.	Subsection 5.2.1
2	Mean length of traditional abstracts is $2L^a$.	Subsection 5.2.1
3	Mean length of structured abstracts is $4L^a$.	Subsection 5.2.1
4	Mean sentence length did not vary significantly by discipline.	Subsection 5.2.2
5	INTRODUCTION MOVE did not occur in one discipline (MED).	Subsection 5.3.2
6	PURPOSE, METHOD, RESULT and DISCUSSION MOVE occurred in all disciplines.	Subsection 5.3.2
7	Abstracts with one move have one potential permutation.	Subsection 5.5.3
8	Abstracts with two moves have two potential permutations.	Subsection 5.5.3
9	Abstracts with three moves have eight potential permutations.	Subsection 5.5.3
10	The number of potential move permutations far exceeded actualized permutations.	Subsection 5.5.3
11	Pairs of moves were repeated in BOT, EC, IND, IP, IT, KDE and WC.[b]	Subsection 5.5.4
12	Trios of moves were repeated in BOT.[c]	Subsection 5.5.4
13	There is one high frequency permutation in the medical sub-corpus.[d]	Subsection 5.5.5
14	There are many high frequency permutations in different disciplines.	Subsection 5.5.5
15	Slightly under 200 move permutations were found.	Subsection 5.5.5
16	The lowest number of moves per abstract was one.	Subsection 5.6.2
17	The highest number of moves per abstract was ten.	Subsection 5.6.2
18	The most frequent number of moves per abstract was four.	Subsection 5.6.2
19	Move sequence is usually linear in MED, MAT, LING and BOT.	Subsection 5.6.3
20	Move sequence is either linear or non-linear in EC, IND, IP, IT, KDE, and WC.	Subsection 5.6.4
21	IMRD[e] and IPMRD[f] were not frequently used.	Subsection 5.6.4
22	Disciplinary corpora can be described using dimensions.	Subsection 5.7.2
23	Three dimensions are linearity, cyclicity and variation.	Subsection 5.7.2
24	A single-digit binary value represents each dimension.	Subsection 5.7.3
25	There are eight regions in and around *Borromean rings*.	Subsection 5.7.3
26	Grammatical tense follows a Zipfian-like distribution.	Subsection 6.2.4
27	68% of all grammatical tenses are *present simple*.	Subsection 6.2.4
28	22% of all grammatical tenses are *past simple*.	Subsection 6.2.4
29	3% of all grammatical tenses are *future simple*.	Subsection 6.2.4
30	Discipline-specific vocabulary[g] is shared across all moves.	Subsection 6.3.2
31	Move-specific vocabulary[g] is *not* shared across all disciplines.	Subsection 6.3.2

[a] L = length, judged by number of word tokens
[b] For example, the sequence MRMR occurred in EC 67, IND 012, IP 18, IT 73, KDE 51 and WC 12.
[c] For example, BOT 013 included MRDMRD while BOT 027 included IMRIMR.
[d] Almost every abstract in MED followed the permutation PMRD.
[e] IMRD was ranked the 15th most frequent (n = 14).
[f] IPMRD was ranked the 16th most frequent (n = 12).
[g] Key words were used as a proxy for vocabulary.

The arguments are presented in an easy-to-follow format enabling total transparency. Each argument is built upon the propositions given in Table 7.1. The arguments are formulated using natural language rather than represented symbolically. Natural language unlike mathematical notation is inherently ambiguous. For

example, even a simple conjunction, such as *or* has two meanings namely inclusive and exclusive. These arguments are presented in the context of this research and so should be read with that in mind rather than being treated as stand-alone decontextualised statements. In some premises the proposition used is categorical, i.e. the grammatical subject is undistributed, e.g. *Some X is Y*. Conclusions based on such propositions are necessarily also undistributed.

Nine claims and their logical arguments are presented in the following subsections. By convention, premises on which the claims or conclusions are based are placed above the horizontal line or inference bar. The conclusion is given below the inference bar. The claims presented are true for the corpus used in this study, but no claim is made that these conclusions hold true when applied to other corpora.

7.4.2 Claim 1: Type of abstract impacts length.

The first claim is that the type of abstract impacts the length of research abstract. Length in this claim is judged by the number of word tokens. Corpus evidence showed that the mean length of structured abstracts was double that of traditional abstracts which, in turn, was double that of graphical abstracts. From this, it can be judged that abstract type and length correlate. This can be represented as a logical argument as follows:

Mean length of graphical abstracts is L
Mean length of traditional abstracts is $2L$
Mean length of structured abstracts is $4L$
Thus, type of abstract correlates to length.

Changes in length of an abstract are achieved by increasing or decreasing the number of word tokens. An increase or decrease in word tokens does not convert a traditional abstract into either a structured or graphical abstract. Changes in the type of abstract involve the addition or subtraction of additional elements. For example, to convert a traditional abstract to graphical abstract, an image is needed. The assignment of causality can be deduced using a disjunctive syllogism as follows:

Either the length impacts the type or the type impact the length.
The length cannot impact the type.
Therefore, type of abstract impacts the length.

The underlying assumption in this argument is that there is no confounding variable affecting causality.

7.4.3 Claim 2: Potential permutations increase exponentially with additional moves.

A research abstract that has five discrete rhetorical moves can be arranged in different 120 permutations when all moves are used and no moves are repeated. An abstract

that has only two moves can only be arranged in two different ways. This potential for multiple permutations is determined by the number of moves.

> Abstracts with one move have one potential permutation.
> Abstracts with two moves have two potential permutations.
> Abstracts with three moves have eight potential permutations.
> Potential permutations increase exponentially with additional moves.

This conclusion can be proven mathematically using Equation 5.1. Given that potential permutations increase exponentially with additional moves, the rhetorical organization is impacted. Abstracts comprising more moves have the potential for a wider variety of permutations. This is supported in the corpus evidence by the lack of repetition of identical move permutations comprising over six moves. It should also be noted that allowing move cycling, repetition of moves and omission of moves further increase the number of potential permutations.

7.4.4 Claim 3: Most disciplines contain linear move sequences.

The argument supporting the claim that rhetorical moves show linearity is given below. Both propositions are categorical involving particular rather than universal case. Using propositions 19 and 20 from Table 7.1 as premises, a conclusion can be inferred as shown below.

> Move sequence is usually linear in MED, MAT, LING and BOT.
> Move sequence is either linear or non-linear in EC, IND, IP, IT, KDE, and WC.
> All disciplines contain linear move sequences.

This claim is based on deductive reasoning for this corpus and can proven by mapping the disciplines showing linearity onto a Venn diagram. However, given the lack of statistical support and a reticence to state that this holds true in every case, a more tentative claim is appropriate: Most disciplines contain linear move sequences. Even for this hedged claim, empirical research is needed to verify its accuracy if it is applied to other disciplines. In the unlikely case that there are a number of disciplines in which no linear abstracts are present, the claim is defeasible.

7.4.5 Claim 4: Some disciplines contain non-linear move sequences.

Using the same premises as in the argument for claim 3, but focussing on non-linearity, another conclusion can be inferred. This claim can also be proven by mapping the disciplines showing non-linearity onto a Venn diagram. However, this time, instances of non-linearity will be mapped to only some disciplines, showing that all disciplines do not contain instances non-linearity.

> Move sequence is usually linear in MED, MAT, LING and BOT.
> Move sequence is either linear or non-linear in EC, IND, IP, IT, KDE, and WC.
> Some disciplines contain non-linear move sequences.

Non-linear abstracts were created by two main ways. First, some abstracts made use of fronting, such as by placing the RESULT MOVE before the METHOD MOVE at the beginning of an abstract. Second, some abstracts cycled through moves, which caused a non-linear juncture between the pair or trio of cycled moves.

7.4.6 Claim 5: Move cycling occurs in some disciplines.

The argument below shows the argument supporting the claim that move cycling occurs in some abstracts. Two examples of move cycling are found in the corpus: one for two moves and one for three moves. Thus, the phenomena of move cycling occurs. The argument can be formulated as follows:

> Pairs of moves were repeated in BOT, EC, IND, IP, IT, KDE and WC.
> Trios of moves were repeated in BOT.
> Move cycling occurs in some disciplines.

Move cycling is one of the causes of non-linearity. For example, in the permutation IMRMR, the adjacent pair of moves MR is repeated twice. By comparing abstracts with move cycling to their accompanying research article, move cycling was found to be used most commonly with the METHOD MOVE and RESULT MOVE when researchers developed an artefact (e.g. algorithm) and then evaluated the performance of the artefact (e.g.speed). However, in the BOT sub-corpus move cycling occurred mainly with the RESULT MOVE and DISCUSSION MOVE. This was because multiple results were reported and each result was discussed in turn.

7.4.7 Claim 6: Move permutations vary by discipline.

The argument to reach the conclusion that move permutations vary by discipline is as follows.

> There is one high frequency permutation in the medical discipline.
> There are many high frequency permutations in different disciplines.
> Therefore, move permutations vary by discipline.

The claim that move sequences vary by discipline is supported throughout the corpus. The most notable differences in variation are between the linear abstracts of MED and abstracts for disciplines, such as IT, IND and WC. The structured abstracts using in MED displayed almost no variation while the engineering (IND) and information science disciplines (EC, IP, IT, KDE and WC) showed a great deal of variation in terms of actualised permutations in the corpus. High degrees of variation occurred in disciplines reporting development and evaluation. These disciplines used both linear and non-linear abstracts, which included move cycling and fronting of moves.

7.4.8 Claim 7: All disciplines can be mapped onto the *Borromean Rings* framework.

The argument showing the logical steps in the creation of the *Borromean Rings* framework is given below. The main claim is that abstracts for all disciplines can be mapped onto the *Borromean Rings* framework. Although this has been tested on the ten disciplines in the corpus, the framework was designed to focus on dimensionality rather than disciplines, and so should be applicable to scientific disciplines not represented in the corpus. The first step was to link dimensionality and disciplines via the syllogistic argument as follows:

> Disciplinary corpora can be described using dimensions.
> Three dimensions are linearity, cyclicity and variation.
> Disciplinary corpora can be described using linearity, cyclicity and variation.

To understand how the dimensionality relates to disciplines, it is necessary to understand the total number of categories and how those categories can be identified. The method selected was to use binary numbers to represent each dimension. Thus,

> A single-digit binary number represents each dimension.
> There are three dimensions.
> A three-digit binary number represents three dimensions.

There are eight possible values [000, 001, 010, 011, 100, 101, 110, 111] for a three-digit binary number and so any framework that has eight regions may be utilized. There are eight regions in and around the Borromean Rings as shown in Subsection 5.7.3. Therefore, it is possible to map each permutation of the three dimensions of linearity, cyclicity and variation onto the *Borromean Rings* framework.

A potential criticism for this framework is the use of dichotomous variables for each dimension. The use of the *Borromean Rings* framework does not, however, rule out adding another feature, such as colour saturation to visualize the magnitude of a variable if a continuous variable is used in place of a dichotomous variable. However, the underlying motivation for the framework is pragmatic and pedagogic and so dichotomous values were selected for ease of use and ease of interpretation. The cut-off point between each of the variables needs to be estimated based on the specific datasets used. One anomaly or one outlier may be safely ignored for pedagogic purposes, but deciding the exact cut-off point of when an anomaly is no longer an anomaly is akin to the paradox of the heap.

7.4.9 Claim 8: Ninety-three percent of all grammatical tenses are *simple* tenses.

Given that the distribution of tense is Zipfian-like, a few grammatical tenses dominate. Thus, by considering using the three most frequent grammatical tenses first, novice-writers can maximize the likelihood that their tense selection is efficient.

68% of all grammatical tenses are *present simple*.
22% of all grammatical tenses are *past simple*.
3% of all grammatical tenses are *future simple*.
93% of all grammatical tenses are *simple* tenses.

Tense carries grammatical meaning and so switching tense may switch meaning. However, context is key. The claim that 93% of all grammatical tenses are *simple* tenses helps focus writers on considering *simple* tenses before considering *perfect* and *progressive* forms. However, at times, the rarer forms, such as *present perfect progressive*, may be more appropriate. It should also be remembered that in the MED corpus a number of sentences in the PURPOSE MOVE did not contain finite verbs.

7.4.10 Claim 9: Discipline-specific vocabulary is more pervasive than move-specific vocabulary.

The argument for the claim that discipline-specific vocabulary is more pervasive than move-specific vocabulary is given below.

Move-specific vocabulary is not shared across all disciplines.
Discipline-specific vocabulary is shared across all moves.
Thus, discipline-specific vocabulary is more pervasive than move-specific vocabulary.

It is unsurprising that the discipline-specific vocabulary occurs throughout texts describing research in the discipline. What is slightly surprising is the lack of shared vocabulary across moves. Taking the METHOD MOVE as a case in point, disciplines such as MED, LING and IND use completely different methods. Axiomatically, different methods are less likely to share the same lexical sets. Experimental studies using human subjects may occur in MED and LING. However, in MED the human subject receives treatment while in LING the human subject may be the source for the language that is analyzed. In IND an electronic circuit may be created and the performance of a particular parameter tested.

In universities that offer research writing tuition to novice writers, students from different disciplines tend to attend the same course. Teachers therefore aim to seek out similarities and may focus on functional exponents for particular rhetorical moves. However, given the knowledge that move-specific vocabulary differs greatly, teachers should avoid overgeneralizing. Ideally, teachers should have direct disciplinary knowledge. Alternatively, the teacher or student can access a corpus to discover the move-specific vocabulary appropriate for a particular discipline.

7.5 Research question 1

The deepest sin against the human mind is to believe things without evidence.
— Thomas Henry Huxley, an English biologist (1825–1895)

The first research question (2.8.2), namely: "What is the rhetorical organization of abstracts of research articles published in a broad range of top-tier scientific journals?" is operationalized into six discrete sub-research questions described in each of the following subsections. The combined answers describe the rhetorical organization.

7.5.1 Sub-research question 1

What moves occur in research abstracts in each discipline?

Based on the literature review, a set of five moves was selected. As described in Section 5.3, of the five moves only the INTRODUCTION MOVE was the only rhetorical move that was not present in all ten disciplines. The remaining moves –PURPOSE MOVE, METHOD MOVE, RESULT MOVE and DISCUSSION MOVE – were used in all ten disciplines. Disciplinary conventions therefore affect the types of moves used.

7.5.2 Sub-research question 2

How frequent is each move in research abstracts in each discipline?

Section 5.4 concludes that the frequency of moves is normally distributed across both the whole corpus and within each discipline in the corpus. The frequency rank however varies by discipline. The value that the community of practice places on particular aspects of research no doubt affect this. In disciplines that value methods, the METHOD MOVE is likely to be found more frequently.

7.5.3 Sub-research question 3

In what sequence do the moves occur in research abstracts in each discipline?

At the level of adjacent pair of rhetorical moves, all twenty potential adjacent pairs of rhetorical moves were found (Section 5.4). Three disciplines – BOT, MAT and MED – used more limited sets of adjacent pairs. The disciplines in information science (e.g. EC, IND, IP, IT, KDE, WC) and industrial electronics (IND) showed the most variation with more than 15 different adjacent pairs per discipline.

At the level of whole abstract, nearly two hundred (n = 196) different permutations of rhetorical moves were discovered. Move sequences comprised between one and ten moves.

Although millions of permutations are possible, just under two hundred were realized. For the MED sub-corpus, the rigid prescribed PMRD structure constrained the number of move permutations while for the MAT sub-corpus which used graphical abstracts, the length limited the number of potential permutations.

7.5.4 Sub-research question 4

How frequent is each sequence in research abstracts in each discipline?

The majority of the frequency distribution were governed by a minority of the permutations. At the level of adjacent pairs of moves, the most frequent three

adjacent pairs of moves accounted for slightly over 50% of all the adjacent pairs. These adjacent pairs were PURPOSE-METHOD MOVES, METHOD-RESULT MOVES and RESULT-DISCUSSION MOVES (Subsection 5.6.3).

At the level of move sequence for the whole abstract, it was found that 20 permutations out of almost 200 accounted for over two-thirds of all the move sequences. The most common permutation in the corpus was PMRD (n = 110) with almost all instances occurring in MED. The next four most frequent permutations were IRMR, PMR, IR and IRD. The sixth most common permutations was a single RESULT MOVE, R, (n = 50), which was particularly common in MAT. The seventh most frequent move involved the fronting of and repetition of a RESULT MOVE, RMR. The commonly advocated move structures of IMRD and IPMRD, however, were not well represented in this corpus and were ranked 15th and 16th most frequent with a combined total of 26 instances.

7.5.5 Sub-research question 5

What are the similarities in rhetorical organization between the disciplines?

The first level of similarities is dictated by the type of abstract, since type constrains the rhetorical organization. Graphical abstracts cannot have multiple moves due to word length limitations while structured abstracts are less likely to have non-linear move sequences. Disciplines which focus on the development and evaluation of artefacts tend to harness the cyclicity dimension.

7.5.6 Sub-research question 6

What are the differences in rhetorical organization between the disciplines?

Differences were found in the dimensions of linearity, cyclicity and variation. Some abstracts were linear (e.g. BOT and MED) while other were not. Some abstracts were cyclical (e.g. EC, IND, IP, IT, KDE and WC) while others were not. When evaluated at the level of disciplinary corpus, different degrees of variation in move sequences could be noticed. Some disciplines (e.g. MAT and MED) showed little variation while other varied greatly.

A notable difference, worthy of further investigation, is the choice of initial move. The first rhetorical move in an abstract acts as a hook. Abstracts beginning with the introduction move were most frequent overall, but the frequency varied by discipline. For example, abstracts in MED begin with a PURPOSE MOVE while those in MAT tended to begin with a RESULT MOVE.

7.6 Research question 2

What are the lexical features of prototypical moves in abstracts of research articles in the selected scientific disciplines?

Adopting the same approach as for the first research question, the combined answers describe the lexical realization. Other researchers have focused on the comparatively straight-forward analysis of key words by generating key word lists for various n-grams and discussing those. In this project, the focus is on the grammatical-end of lexical realization.

7.6.1 Sub-research question 7

Does the lexical realization differ between the same moves in different disciplines?

Subsection 6.2.3 and 6.2.4 have shown that lexical realization judged by keyness and grammatical tenses differed in the same move across different disciplines. The top five key words in each move also differed for each discipline. There were more similarities among rhetorical moves in the same discipline than among the same moves in differ disciplines.

In terms of grammatical tenses, almost all of the grammatical tenses were *simple* tenses. Although *simple* tenses dominate the whole corpus, when tense usage was compared among the disciplines, there were many commonalities; yet the spread of tenses and the order of the top ranked tenses varied (See Subsection 6.2.4).

7.6.2 Sub-research question 8

Does the lexical realization differ between different moves in the same discipline?

As shown in 6.3.2 and 6.3.3, the lexical realization measured in terms of keyness and grammatical tenses differed between different moves in the same discipline.

There was more similarity between moves in the same discipline than across the same move in different disciplines. Many terms within the same discipline occurred in the top five most frequent words in each move. However, the rank frequency of the top five key words differed in almost every move.

7.6.3 Sub-research question 9

To what extent does the lexical realization differ between moves?

Lexical realization differed by move. When rhetorical moves were plotted in the same vector spaces, the moves were distinct and separate regardless of features selected. This showed that lexical realization judged by keyness and grammatical tense differs between rhetorical moves. Although *present simple* form is the most frequent grammatical tense, the rank frequency and distribution frequency of grammatical tenses varied across rhetorical moves.

7.6.4 Sub-research question 10

To what extent does the lexical realization differ between disciplines?

The lexical realization differed by discipline. When the ranks of key words and grammatical tenses were compared among the same rhetorical move in different

disciplines, there was little similarity. Greater degrees of similarity were found between rhetorical moves in the same discipline than between the same rhetorical move in different disciplines. Simply put, discipline-specific lexical choice appears to trump move-specific lexical choice.

7.7 Pedagogic implications

Knowledge is of no value unless you put it into practice.
— Anton Chekhov, Russian playwright

7.7.1 Genre-based approach

This study has shown that despite the extensive variation in rhetorical organization among the disciplines, there are discernible patterns. Knowledge of the patterns of rhetorical organization and lexical realization can be harnessed by both teachers and novice writers. Teachers of both English for Specific Purposes and English for Research Publication Purposes often state the necessity to check specialist corpora for the target genres to understand the expectations of the community of practice. This advice is valid. However, with knowledge of the possible pattern to look for, scientific abstracts that have not been described in the research literature may be analyzed quickly using the *Borromean Rings* framework.

TABLE 7.2: List of claims based on arguments with premises taken from propositions supported by corpus evidence

Number	Claims	Reference
1	Type of abstract impacts length.	Subsection 7.4.2
2	Potential permutations increase exponentially with additional moves.	Subsection 7.4.3
3	Most disciplines contain linear move sequences.	Subsection 7.4.4
4	Some disciplines contain non-linear move sequences.	Subsection 7.4.5
5	Move cycling occurs in some disciplines.	Subsection 7.4.6
6	Move permutations vary by discipline.	Subsection 7.4.7
7	All disciplines can be mapped onto the *Borromean Rings* framework.	Subsection 7.4.8
8	93% of all grammatical tenses are *simple* tenses.	Subsection 7.4.9
9	Discipline-specific vocabulary is more pervasive than move-specific vocabulary.	Subsection 7.4.10

Table 7.2 shows the list of the claims that can be made based on propositions related to both rhetorical organization and lexical realization. The arguments for claims related to rhetorical organization and lexical realization are given in Section 7.4. Front-line teachers of research writing can use these claims to inform their practice. Many teachers of English may have backgrounds in the humanities rather than applied sciences and so be unaware of the degree of variation that is present in such abstracts. The concepts of cycling through moves and fronting of moves are also likely to be unknown. Teachers of research writing who have read some of the literature on research abstracts are likely to be aware of IMRD and CARS, and mistakenly assume that all scientific disciplines adhere to these models.

Disciplinary conventions

Given the radical differences in types of abstracts, it is vital to understand the type of abstract that is appropriate for the publication venue. This is dependent on both the disciplinary conventions and the editorial preferences. Once the type of abstract has been identified, the generic expectations of the community of practice need to be understood. Mentorship from an experienced practitioner in the field is the ideal way to make the journey from the periphery to the core of the discourse community. It is necessary to understand what is valued by community members. In terms of scientific abstracts, it is useful to know the amount of value that is placed on novelty, significance, substance and rigour. Generic integrity can only be achieved if writers have sufficient generic awareness of the language features utilized.

Deconstruction, reconstruction and creation

One way to increase genre awareness is to use the approach advocated by the Sydney School of Systemic Functional Linguistics. In this approach learners are engaged in collaborative activities that focus on both genre and language use. The collaborative approach necessitates discussion and ensures that learner are actively engaged in the three core activities of deconstructing texts, reconstructing texts and creating their own texts. Typical activities could involve jigsaw reading in which students read a small part of a larger text and then collaborate together to complete a task, such as drafting an abstract for the text. Sentence skeletons, sentence stems and guiding questions can be used to help learners make their initial drafts.

According to the noticing hypothesis, learners need to notice a language feature before being able to learn the feature. Therefore, any activity that engages learners in noticing can be considered to have pedagogic value.

Data-driven

Due to the degree of disciplinary variation, it is essential for writers to understand the disciplinary conventions in their own field. This sense for what is used and not used is often acquired through extensive exposure. However, for researchers primarily operating in and reading text written in languages other than English, it is necessary to learn about the generic conventions is intensively.

Teachers of writing with sufficient disciplinary knowledge may be able to provide novice writers with advice; but should that not be possible, data needs to be obtained. Depending on the context either a corpus-based approach or a data-driven learning approach could be used. in both approaches novice writers can focus on language features in a specialized corpus or data set. As the purpose is pedagogic rather than rigorous research, a corpus of just 5 to 10 is likely to be sufficient to enable both teachers and students to understand the expected conventions.

7.7.2 Implications for teaching rhetorical organization

Disjuncture

There is a paucity of details on the rhetorical structure of scientific research abstracts. Currently published pedagogic materials for scientific writing provide little guidance on the many different types of research abstracts. For example, no textbooks were found that mentioned the wide-spread use of structured abstracts or graphical abstracts. Although the IMRD structure is well described, the numerous other permutations receive almost no mention. Neither the high frequency of move cycling in some domains, (e.g. wireless communication) and the fronting of moves, particularly in abstracts of short articles, such as letters or communications, were mentioned either. Teachers of research writing who specialize in assisting writers in particular disciplines most likely soon discover the disjuncture between the pedagogic materials and published abstracts; but teachers who help groups of writers from a wide variety of research disciplines may remain unaware of the vast differences among the disciplines without knowing this information or discovering it themselves via their own corpus investigation. Using the results of this study, teachers of research writing can develop teaching materials that are research-supported.

The advice disseminated in most research writing text based advocates the use of IMRD or IPMRD. Yet, this study showed that there were 194 other permutations of rhetorical moves and their neither IMRD nor IPMRD were frequently used. Providing learners with an initial model to introduce the main concepts is a valid pedagogic strategy, but learners need to be aware of the whole forest of permutations rather than just the IMRD and IPMRD trees. There is a disjuncture between the prescriptive advice and the descriptive reality. The plethora of permutations are pared down to one pair. Data-driven learning and/or genre-based analysis could be used to show learners the range of permutations in their own subject disciplines.

Dimensionality

Once learners are aware of the type of research abstract, e.g. graphical, structured or traditional; the next consideration is to understand the value that the community of practice places on novelty, significance, substance and rigour. Armed with this information, students can focus on rhetorical organization.

Based on a study of evaluation in 300 research abstracts Stotesbury (2003) concluded that it was not useful to teach students the rhetorical structure of abstracts as suggested by the general guidelines because each discipline follows its own convention of abstract writing. Hyland (2004, p.3) explained that disciplinary differences manifest themselves in different ways of truth construction and reader engagement.

Thus, rather than starting with move-based genre analysis, students could first be introduced to the three dimensions of linearity, cyclicity and variation. This enables learners to take a more holistic view of research abstracts and focus on patterns rather

than individual moves. Understanding the common patterns provides learners with a schema onto which the specific types of move permutations can be placed.

The dimensions of linearity, cyclicity and variation could be introduced and practiced without any text. Possible patterns of abstracts can be introduced using coloured cards that represent different moves (e.g. main result, secondary result) or groups of moves (development phase, evaluation phase). This activity provides teachers with the opportunity to show learners how moves may be fronted for emphasis in some disciplines, and how pairs of adjacent moves may be used to describe the development and evaluation of artefacts. This activity enables students to share their disciplinary knowledge. This can then lead into working with actual abstracts.

Demarcation

Teachers of scientific research abstracts who understand the rhetorical territories and borders of different disciplines are better placed to assist learners. The *Borromean Rings* framework can help organize disciplines according to the rhetorical dimensions. For teachers working with larger classes, this may enable them to match students according to the regions within the framework.

7.7.3 Implications for teaching lexical realization

Before considering lexical realization, learners need to know:

1. type of abstract (e.g. traditional);
2. community value placed on novelty, significance, substance and rigour;
3. permutations commonly used in their discipline; and
4. the rhetorical structure to use for their abstract.

It might be that some novice writers draft their abstract in a language other than English first and then use software to translate the abstract. Using this translation, the final abstract is constructed. With the increase in the accuracy of machine translation, this is a trend that is likely to increase. To enable learners to make the appropriate choices of vocabulary and grammar, knowledge of collocation and colligation can help.

Collocation

A useful item for students who need to grasp technical language quickly is to focus on the content words that are used disproportionately frequently within a discipline. Lists of the top 10, top 25 and top 100 key words could be used to help learners get to grips with the commonly used terms. Initially, single key word lists could be used. This could be followed by lists of meaningful key bigrams and trigrams.

Students could also be introduced to the Academic Phrasebank (Morley, 2004; Morley, 2020), which is a database of usable functional exponents for academic writers. Although the Phrasebank was not developed specifically for scientific research abstracts, it is a valuable resource that can help writers struggling with vocabulary selection.

Colligation

There is a correlation between grammatical tense and rhetorical moves used in each discipline. Overall, the pattern was that *present simple* was the most frequent tense in almost all moves in all disciplines. There were some exceptions, however. This implies that teachers and students need to check the disciplinary conventions before finalizing an abstract. In the MED corpus, for example, *past simple* tense was the most commonly used tense in the RESULT MOVE. As *simple* tenses accounted for almost all (96%) of the grammatical tenses found in the corpus, novice writers should concentrate on these three grammatical tenses.

Another useful resource is the Academic Word Suggestion Machine (Mizumoto, Hamatani, and Imao, 2017), which is an online tool that helps writers complete sentences. This tool focuses on the colligation of lexical bundles and rhetorical moves. Users can select the particular genre (e.g. abstract) and a subject discipline (e.g. computer science), after which when writers type, the software generates predictive text based on word bundles.

Default, decision tree and detailed drafts

As described in detail in Subsection 6.6.4, one way in which teachers can help students draft abstracts in a timely manner, yet select the most appropriate grammatical tenses is to adopt a cyclical approach. In this way, students create multiple drafts (or revisions) or a research abstract with each subsequent draft increasing in grammatical sophistication. In the first draft all verbs are simply assigned the default tense of *present simple*. Statistically, 70% of the verbs will be assigned the most appropriate tense. In the second draft, the decision tree shown in Figure 6.21 is used to check each of the verbs. This creates a draft in which three simple tenses have been considered for each verb. In the third and subsequent drafts, all grammatical tenses can be considered. Writers with higher levels of grammatical competence may switch some of the simple tenses to more appropriate tenses while those without the ability to distinguish between the finer nuances should still have a draft that is sufficiently accurate.

7.8 Limitations

Real knowledge is to know the extent of one's ignorance.
— Confucius

Through the use of corpus linguistics it was possible to identify novel and hitherto hidden patterns of rhetorical organization and lexical realization. As with all approaches, there were a number of limitations. However, as Fillmore (1992, p.35) notes facts about the nature of language would remain undiscovered if a corpus had not been utilized, and so a corpus is needed regardless of any possible deficiencies regarding size and so forth.

To quote Sinclair (2004c, p.10),

> It is impossible to study patterned data without some theory, however primitive. The advantage of a robust and popular theory is that it is well tried against previous evidence and offers a quick route to sophisticated observation and insight. The main disadvantage is that, by prioritizing some patterns, it obscures others.

Viewing the data through the lens of rhetorical moves enabled previously unpublished patterns of rhetorical moves to be discovered, but simultaneously the opportunity to discover hitherto hidden patterns unrelated to rhetorical moves was lost.

Three limitations regarding the corpus and annotation accuracy are given below.

The first limitation is related to the size of the corpus of scientific research abstracts. The size of the corpus and the accuracy of the annotations are easily sufficient for the purposes of this research study. However, should a more granular study be conducted, it would be necessary to have a larger corpus. A case in point is that when the fifty sub-corpora of abstracts grouped by both move and discipline were created, some of the sub-corpora that were too small for some types of statistical analysis. For example, cluster analysis carried out on insufficiently large datasets will produce unreliable results, which is why no investigation of the similarity between move-disciplines could be undertaken using cluster analysis.

The second limitation relates to the quality of annotations. In an ideal world where money and time were infinite resources, more annotators who were experts in the disciplines would be used to annotate larger quantities of the corpus. This would help to counter any criticism of annotator accuracy. However, in reality the cost of annotation is high, and so without sufficient resources, pragmatic decisions need to be made to ensure project completion within the time and financial constraints. This is not to say that the accuracy would necessarily be improved by further annotation, just that the inaccuracies would be spread over multiple annotators, and so bias would be spread. Should the annotators of the primary investigator be more accurate than the mean annotation accuracy rate of a pool of annotators, the overall accuracy would, in fact, drop.

The third limitation of this study is the accuracy of the automatic tense identification. The corpus notionally contains 7200 sentences. The classification as a sentence was based on the presence of a capital letter, a meaningful unit or message and the presence of an end stop. A minority of these sentences, however, do not contain

finite verbs. This was frequently the case in the PURPOSE MOVE of MED abstracts. The majority of sentences in the corpus contained one finite verb, but where subordination or coordination occur two or more tenses may be identified. In total 6753 sentences were assigned grammatical tenses in the corpus. Tense was not discovered in at least 447 sentences either due to the lack of a finite verb or an identification error. Given the purpose of comparing the moves and disciplines together, this error rate should not affect the analysis since false positive and false negative results will occur throughout the corpus and should not significantly affect one discipline or move more than another. This is also true for almost all automated methods. For example, automatic annotation of texts with POS-tags is reported to be approximately 97% accurate (dependent on the tagger), and a three percent error is perceived as acceptable. The reality, however, is that the 3% error is per word and so the mean accuracy for the mean sentence length in this corpus of 23.5 words drops drastically to slightly under 49% for POS-tagging accuracy.

7.9 Future work

One area in which future work is deserving is in the automated classification of rhetorical moves. This is a research field that is more in the realm of computational linguistics rather the corpus or applied linguistics. However, should a sufficiently accurate automated classifier be available, applied linguistics would be able investigate the rhetorical organization of research abstracts without incurring such a large time cost.

Automated identification of grammatical tenses is another area in which future work is needed. This classification task is easier than identification of rhetorical moves since grammatical tenses can be identified based on syntactic rather than pragmatic structure. The task, however, is still non-trivial. A central difficulty is that tense identification is contingent upon the accuracy of part-of-speech (POS) tagging. However, if there are no breakthroughs in the accuracy of POS tagging, it may be possible to post-process the results to increase the accuracy of any tense identification program.

This research focussed on describing a corpus of scientific research abstracts. However, the applicability of this description to a wider population is unclear. Although it is likely that the claims may apply, no experimental research was conducted. This could be a fruitful area of research.

One feature that was not considered in this research project was the specific type of research article that the abstract accompanied. Research articles may vary not only with discipline, but according to type of article. For example, the rhetorical organization of research abstracts for full-length research articles and short research articles may differ. Abstracts of experimental, theoretical and empirical research articles may also differ. An additional round of annotation could be undertaken

to investigate any correlation between rhetorical organization and type of research article.

The latest form of abstracts, video abstracts (Bredbenner and Simon, 2019), was not represented in the corpus. Given an apparent online trend to multimodal resources on the internet in general, it appears likely that in order to make research more accessible, video abstracts or summaries may be adopted. The first video abstract was published in the journal *Cell* (Enard and Svante, 2009). *Cell* was one of the first journals to adopt graphical abstracts; and in addition to graphical abstracts regularly publishes three-minute video abstracts. However, unlike written abstracts which may be drafted rather quickly, video abstracts take substantially longer to produce. The pay-off, however, is in the increased interest in and dissemination of the research results. An investigation into the rhetorical moves in video abstracts, graphical abstracts (image and text) and textual abstracts could provide novel insights.

Appendix A

Appendices

A.1 Algebraic representation of research problem

The following is a first-principles approach to representing the research problem algebraically.

Let $\mathcal{M} = \{I, P, M, R, D\}$ be the set of moves. $x_i^{(j)} \in \mathcal{M}$ is the move in sentence i of document j. If X is one of $\{I, P, M, R, D\}$, then number of moves in document j is:

$$\sum_{i:x_i^{(j)}=X} x_i^{(j)}. \tag{A.1}$$

Alternatively using set-builder notation, the set of indices of sentences with move X is given by:

$$\{i \mid x_i^{(j)} = X\}$$

so that the number of such indices is the cardinality (size) of the set:

$$\#\{i \mid x_i^{(j)} = X, \forall i\} \text{ or } |\{i \mid x_i^{(j)} = X, \forall i\}| \tag{A.2}$$

So that (A.1) and (A.2) are supposed to be the same.

Suppose discipline k consists of documents $\mathcal{D}_k = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots\}$ where $\mathbf{x}^{(j)} = [x_1^{(j)}, x_2^{(j)}, \ldots]$ represents one document. Then the total number of moves in that discipline is denoted by a counting function C:

$$C(\mathcal{D}_k, X) = \sum_{j \in \mathcal{D}_k} \sum_{i:x_i^{(j)}=X} x_i^{(j)} \tag{A.3}$$

Alternatively using set-builder notation, the set of index pairs of (documents, sentences) with move X is given by

$$\#\{(j,i) \mid x_i^{(j)} = X, \forall i, \forall j\} \text{ or } |\{(j,i) \mid x_i^{(j)} = X, \forall i, \forall j\}| \tag{A.4}$$

So that (A.3) and (A.4) are supposed to be the same.

The above counts the number of moves in each sentence, so two moves of the same type next to each other count twice. To avoid this, sentences $\mathbf{x}^{(j)}$ are postprocessed to

remove such duplicates.

With thanks to Brian Kurkoski.

A.2 Annotation guidelines

Overview

Tool

Manual coding is undertaken using UAM Corpus Tool 3.0 (O`Donnell, 2014). The annotations in the Corpus Tool are anchored in the text using angle brackets, < and >. The UAM Corpus Tool is available for free download from www.wagsoft.com/CorpusTool/. This tool uses layers to separate different levels of annotation. The focus of this booklet is on coding the move layer. The terms code, tag, annotate and mark up all refer to the process of assigning a particular code (i.e. name of a move or sub-move) to a tag (i.e. embedded marker) which marks up or annotates (e.g. provides more details for) the original text.

Prior to coding moves, Metadata has been marked up with provenance information in a separate layer of annotation in the UAM Corpus Tool. The metadata consists of the research domain (e.g. Information Theory) and publication name (e.g. Transactions on Information Theory)

Move Tagset

There are five main categories and one temporary category (shown in Table 1). The five main categories are: introduction, purpose, methodology, result/product and discussion/conclusion. The temporary category is a way to mark up sentences that are not easily assigned to one or more of the five categories. The temporary category can be visited later to reassign the sentences to a more suitable category. There are five minor categories, namely background, problem, gap, overview and method as product. The major categories are classed as moves, while the minor categories are sub-moves.

Table 1: Tagset for rhetorical moves

Category	Tags
Introduction	<introduction> </introduction>
Background	<background> </background>
Problem	<problem> </problem>
Gap	<gap> </gap>
Overview	<overview> </overview>
Purpose	<purpose> </purpose>
Methodology	<method> </method>
Result (Product)	<result> </result>
Method as product	<methodasproduct> </methodasproduct>
Result as product	<resultasproduct> </resultasproduct>
Discussion (Conclusion)	<discussion> </discussion >
Uncertain	<uncertain> </uncertain>

Schema

The schema (see screenshot in Figure 1) is a tailormade tag set that specifies the hierarchy of tags to markup text in an efficient manner. Once an abstract is open in the UAM Corpus Tool, the first ontological unit is automatically selected. The annotator then is given a choice from menu items for the first level of hierarchy (MOVE TYPE). If the move is tagged as <introduction> or <result>, a second level of hierarchy then appears (INTRODUCTION TYPE or RESULT TYPE). The annotator then selects from the choice given. Once a sentence has been fully tagged, the next ontological unit is automatically selected.

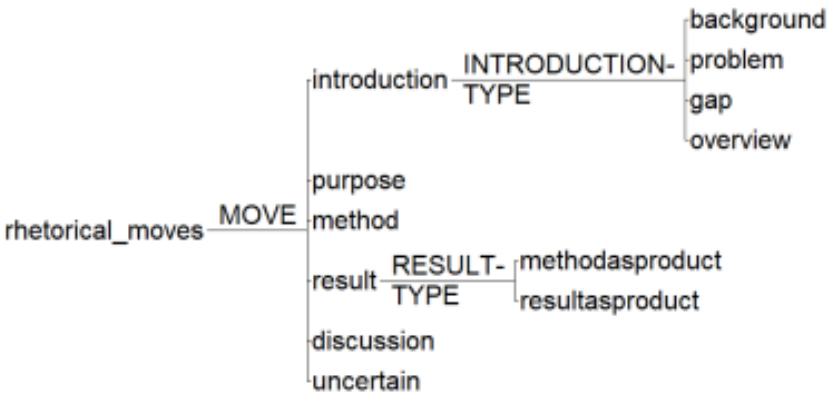


Figure 1: The tailormade schema to manually annotate moves in research abstracts

Procedure

Open UAM Corpus Tool 3.0 and select the corpus to code and the schema.

Ontological unit
The ontological unit is the sentence. The UAM Corpus Tool is configured to automatically identify sentences using regular expressions that search for end stops. Note that when end stops occur within a sentence, automatic identification may fail. In these cases, the annotator needs to manually select the text to be annotated.

Gist reading
Read the complete abstract to get an overall picture. If necessary, look up any terms or concepts needed to understand the organization of the abstract and then re-read the abstract. Remember the aim is not to understand the research itself, but to understand the organization of the abstract.

Specific reading
Read each sentence in order and assign a move and sub-move (for introduction and result moves only). Assign the tag based on your view of the author`s intention. The context and location of the sentence within the text provide clues. If in doubt, assign as uncertain. If the sentence contains two moves, assign one move, then select the sentence again and assign the other move.

Follow-up reading
Reread the complete abstract and (if possible) assign move to those temporarily categorized as uncertain. Check that every sentence in the text has been marked up (tagged).

Save and close the abstract.

Nomenclature

The rhetorical moves are classified based on the nomenclature adopted by Hyland (2004, p.67), which is provided in Table 2. Hyland preferred the term product to describe the move that includes the result which may be a product although in this research, the term result is used. This difference in term does not imply a difference in meaning between Hyland`s definition and mine.

Table 2: A classification of rhetorical moves in article abstracts

Move	Function
Introduction	Establishes context of the paper and motivates the research or discussion.
Purpose	Indicates purpose, thesis or hypothesis, outlines the intention behind the paper.
Method	Provides information on design, procedures, assumptions, approach, data, etc.
Product	States main findings or results, the argument or what was accomplished.
Conclusion	Interprets or extends results beyond scope of paper, draws inferences, points to applications or wider implications.

Source: Table 4.1 in Hyland (2004, p.67)

Hyland (2004, p.67) notes the rationale for separating the purpose from the introduction is due to the role of the writer`s purpose, which may differ greatly from the purpose of providing contextual background for the research.

In order to systematize the manual identification of rhetorical moves, clear definitions for each move are important. However, given the potential subjectivity of a reader trying to identify the author`s intention for each sentence, some guidance is necessary.

Table 3 aims to provide annotators with both the definition of the function for each move, and an overarching guiding question that the move is related to.

Table 3: Functions and guiding questions

Move	Function	Guiding questions
Introduction	Establishes context of the paper and motivates the research or discussion.	Why do this research?
Purpose	Indicates purpose, thesis or hypothesis, outlines the intention behind the paper.	What is the aim?
Method	Provides information on design, procedures, assumptions, approach, data, etc.	How was the research conducted?
Result	States main findings or results, the argument or what was accomplished.	What was found or created?
Conclusion	Interprets or extends results beyond scope of paper, draws inferences, points to applications or wider implications.	So what?

Two moves, namely <introduction> and <result> contain sub-moves. Sub-moves are functional expressions that may combine with other sub-moves. Sub-moves have been namely as steps (Swales, 1990) and stages (Bhatia, 1993) but in order to avoid ascribing to either dichotomy, sub-moves (Bhatia in Hewings, 2006) are used. Sub-moves may be identified prior to moves. By definition, since sub-moves are allocated to particular moves, once a sub-move is identified, the move to which the sub-move belongs is also identified. For example, on reading a sentence, when you identify that the sentence describes a problem and that the problem is one that the research aims to address, then the sentence will be coded as: <introduction> and <problem> . Table 4 provides definitions of functions and guiding questions for the five sub-moves.

Table 4: Functions and guiding questions for sub-moves

Move	Function	Guiding questions
Background	Provides readers with knowledge necessary to follow the research or with details of earlier studies.	What concepts underlie this study? What other studies have been done in this area?
Problem	Describes the problem that the research aims to address.	What problem does this research address?
Gap	Is identified that no other research has yet addressed.	Is this research covering new ground?
Overview	Provides an advance organization of the structure of the article	How is the paper organized?
Result as product	States that the main outcome of the research is not the creation of a method	Is the method the outcome of the result? No.
Method as product	States that the main outcome of the research is the production of a method	Is the method the outcome of the result? Yes.

Further details for the five key moves

Introduction

Definition:	Establishes context of the paper and motivates the research or discussion (Hyland, p.67).
Indicative contents:	background research, background knowledge, problem identification, research gap identified, preview of organization of paper
Key question answered:	Why do this research?

Examples:

1. The paper presents a problem motivated by the hidden subgroup problem, for which the "standard approach" is to use the oracle to produce the coset state. [IT 001]

2. The problem of finding a minimum decomposition of a permutation in terms of transpositions with predetermined non-uniform and non-negative costs is addressed. [IT 002]

3. The reality gap, which often makes controllers evolved in simulation inefficient once transferred onto the physical robot, remains a critical issue in evolutionary robotics (ER).[EC 060]

All three sentences introduce a problem or issue (that can be viewed as a problem) and are therefore tagged as <Introduction> and <Problem>.

4. In the Euclidean optimal communication spanning tree problem, the edges in optimal trees not only have small weights but also point with high probability toward the center of the graph. [abstract EC 008]

This sentence mentions a problem and so is tagged as Introduction. But, from only this sentence it is not clear whether the research addresses the Euclidean optimal communication spanning tree problem or a problem within the problem. The mentioning of Euclidean optimal communication spanning tree problem could simply be background information to the actual problem addressed. On further reading, it is confirmed that the research does not address the Euclidean optimal communication spanning tree

problem and so the sentence is tagged as <Introduction> and <Background>. The keyword "problem" leads the annotator to first identify "problem identification" as a sub-move, but it is necessary to read all the contextual information before assigning a move.

5. In this paper, we describe a study of the evolution of consensus, a cooperative behavior in which members in both homogeneous and heterogeneous groups, must agree on information sensed in their environment.[EC 078]

This sentences introduces the study and provides some general information for lay readers, namely the need for agreement of group members, and is therefore tagged as <Introduction> and <Background>.

Key problem

Purpose in this framework is a move that stands at the same hierarchical level of introduction. The purpose itself could be viewed as the motivation or context for the research. However, code purpose statements only as <Purpose>. Do not double code purpose statements as both <Introduction> and <Purpose>.

Potential indicators of introduction move

1. Sentences whose subject is "this research/paper/study" with finite verbs in present simple tense, such as "presents/shows/addresses"
2. Sentences that introduce a problem which the research addresses.

Purpose

Definition:	Indicates purpose, thesis or hypothesis, outlines the intention behind the paper (Hyland, p.67).
Indicative contents:	purpose aim objective research question hypothesis
Key question answered:	What is the aim?

Examples:

1. The question is: How many copies of the unknown state does one need to be able to distinguish them all with high reliability? [IT 001]

This sentence states the research question and so is tagged as <Purpose>. Note though that if the question introduced is not the research question, but a more general question providing background information, then code the question as <Introduction> and <Background>.

2. The aim of this paper is to prove coding theorems for the wiretap channel and the secret key agreement based on the the notion of a hash property for an ensemble of functions. [IT 051]

3. The primary objective of this research is to propose and investigate a novel ant colony optimization-based classification rule discovery algorithm and its variants.

These sentences overtly state the aim or objective of the research and so are tagged as <Purpose>.

Method

Definition:	Provides information on design, procedures, assumptions, approach, data, etc. (Hyland, p.67).
Indicative contents:	procedure materials algorithm
Key question answered:	How was the research conducted?

Examples:

1. Abstractly, one is given a set of quantum states on a d-dimensional Hilbert space, with the property that the pairwise fidelities are bounded. [IT 001]

This sentence states the parameters of the underlying mathematical model, and so is coded as <Method>.

2. The minimal state will depend on the precise geometric position of the states relative to each other, but useful bounds can be obtained simply in terms of the number N and the fidelity F.[IT 001]

This sentence also states the parameters of the underlying mathematical model, and so is coded as <Method>.

3. For such cost functions, polynomial-time, constant-approximation decomposition algorithms are described. [IT 002]

This sentence names the algorithms to be used in the procedure, and so is coded as <Method>.

Key problem

Research in information science may focus on the creation of new algorithms. The outcome or product or research is therefore a new algorithm, which is intrinsically a method. Therefore, when coding ensure that when a method is described, decide whether the author intended the method to be a description of how the research was conducted or what the research found. It is possible (but unlikely) for exactly the same sentence to appear as a method in one abstract, but as a result in a different abstract.

Potential indicators of method move

1. Sentences using the word "method"
2. Sentences describing parameters, e.g. boundaries and limits.

Results

Definition:	States main findings or results, the argument or what was accomplished (Hyland, p.67).
Indicative contents:	findings results outcome product
Key question answered:	What was found or created?

Examples:

1. Experiments showed that the proposed object segmentation method outperforms others by 21% in the PRI. [IP 074]

This sentence states the findings and so is tagged as <Result>.

2. Considering a range of real-world and representative gartificialh datasets, we show that the method is able to provide relatively low cost solutions for far larger tables than is possible for the optimal approach to tackle. [EC 027]

This sentence states the product of the research and so is tagged as <Result> and because the product is a method, the sub-move tag <Methodasproduct> is also added.

3. Finally, a proof is provided for the fact that the optimal tradeoff between error exponents of a two-message code does not improve with feedback on discrete memoryless channels (DMCs) [IT 003]

This sentence could be a justification of a method. But, in fact, the research itself was to develop an algorithm and this is the evidence of its success. This sentence is therefore tagged as <Result>.

4. Recombination operators of direct encodings like edge-set and NetDir can be extended such that they prefer not only edges with small distance weights but also edges that point toward the center of the graph. [EC 008]

This sentence could be a description of method (extension of recombination of operators) or the product of the research. Given that the main thrust of this research study is the creation of a new method and proving that method is effective. This is classed as <Results>, <Methodasproduct>. .

Potential indicators of results move

1. Sentences using phrases that assert something has been found or created.

Discussion

Definition:	Interprets or extends results beyond scope of paper, draws inferences, points to applications or wider implications (Hyland, p.67).
Indicative contents:	discussion conclusion inferences applications implications generalizations
Key question answered:	So what?

Examples:

1. The presented algorithms have a myriad of applications in information theory, bioinformatics, and algebra. [IT 002]

This sentence concludes that the research results can be applied to three domains, and so the sentence is coded as <Discussion>.

2. Although this research is specific to credit application fraud detection, the concept of resilience, together with adaptivity and quality data discussed in the paper, are general to the design, implementation, and evaluation of all detection systems.[KDE 038]

This sentence concludes that the research results can be applied not only to the data in the paper, but generally to all detection systems, and so the sentence is coded as <Discussion>.

3. These theorems imply that codes using sparse matrices can achieve the optimal rate.[IT 051]

Implying involves generalizing from particular to universal and so the sentence is coded as <Discussion>.

Potential indicators of discussion move

1. Sentences that include words, such as implications and applications.
2. Sentences that use modality to induce their claims to a wider context.

Annotations

Txt file tagged at sentence level

The txt file of an annotated abstract is shown below:

> <purpose> In this paper, we evaluate the applicability of genetic programming (GP) for the evolution of distributed algorithms. </purpose> <method> We carry out a large-scale experimental study in which we tackle three well-known problems from distributed computing with six different program representations. </method> <method> For this purpose, we first define a simulation environment in which phenomena such as asynchronous computation at changing speed and messages taking over each other, i.e., out-of-order message delivery, occur with high probability. </method> <method> Second, we define extensions and adaptations of established GP approaches (such as tree-based and linear GP) in order to make them suitable for representing distributed algorithms. </method> <method> Third, we introduce novel rule-based GP methods designed especially with the characteristic difficulties of evolving algorithms (such as epistasis) in mind. </method> <discussion> Based on our extensive experimental study of these approaches, we conclude that GP is indeed a viable method for evolving non-trivial, deterministic, non-approximative distributed algorithms. </discussion> <discussion> Furthermore, one of the two rule-based approaches is shown to exhibit superior performance in most of the tasks and thus can be considered as an interesting idea also for other problem domains. </discussion>

Summary annotation

Since moves are tagged at sentence level in order to establish the overall move structure of an abstract, the move boundaries need to be identified. A move boundary is created when the adjacent sentence tags differ. Consider the example show in Table 5 below.

Table 5: A classification of rhetorical

Sentence	Move assigned to each ontological unit	Overall move structure
1	Purpose	Purpose
2	Method	Method
3	Method	

4	Method	
5	Method	
6	Discussion	Discussion
7	Discussion	

Move boundaries occur between sentences 1 and 2, 2 and 3, and 5 and 6. However, there are no move boundaries between sentences 2 and 3, 3 and 4, 4 and 5, and 6 and 7. Although the abstract contains 7 sentences each tagged with a move, the transition from one move to another only occurs three times and so the resultant move structure contains three moves, namely PMD. In short, sentences with the same tag are amalgamated into one ontological unit, namely: a rhetorical move.

Txt file tagged at move level

In preprocessing stage, the annotated abstract can be converted to show the overall move structure using regular expressions. The txt file of the overall move structure of abstract is shown below:

> <purpose> In this paper, we evaluate the applicability of genetic programming (GP) for the evolution of distributed algorithms. </purpose> <method> We carry out a large-scale experimental study in which we tackle three well-known problems from distributed computing with six different program representations. For this purpose, we first define a simulation environment in which phenomena such as asynchronous computation at changing speed and messages taking over each other, i.e., out-of-order message delivery, occur with high probability. Second, we define extensions and adaptations of established GP approaches (such as tree-based and linear GP) in order to make them suitable for representing distributed algorithms. Third, we introduce novel rule-based GP methods designed especially with the characteristic difficulties of evolving algorithms (such as epistasis) in mind. </method> <discussion> Based on our extensive experimental study of these approaches, we conclude that GP is indeed a viable method for evolving non-trivial, deterministic, non-approximative distributed algorithms. Furthermore, one of the two rule-based approaches is shown to exhibit superior performance in most of the tasks and thus can be considered as an interesting idea also for other problem domains. </discussion>

A.3 C script to count move combinations

This script was created to identify the combinations of moves and sub-moves at three levels of granularity, namely:

1. all five moves and sub-moves
2. all five moves but no sub-moves
3. four moves with purpose move downgraded to submove in introduction

```
#include <stdio.h>
#include <stdlib.h>
#include <ctype.h>
#include <string.h>

int main()
{
        FILE *iptr, *optr, *aptr, *bptr;
        int i, j, n = 0, a, p, b, f, q, count = 0, flag = 0;
        char lvlcode[110][50], store[50], store2[30];

        if((iptr = fopen("abstracts.txt", "r")) == NULL)
        {
                printf("Error! File could not be read.\n");
                exit(1);
        }

        optr = fopen("Sentence Level Code.txt", "w");
        aptr = fopen("5-Move Code.txt", "w");
        bptr = fopen("4-Move Code.txt", "w");

        while(!feof(iptr))
        {
                fgets(lvlcode[n], 50, iptr);
                n++;
        }

        for(i = 0; i < n; i++)
        {
                a = 0;
                for(j = 0; lvlcode[i][j] != '\0'; j++)
                {
                        if(isupper(lvlcode[i][j]) != 0)
```

```
34                      {
35                              fprintf(optr, "%c ", lvlcode[i][j]);
36                              store[a] = lvlcode[i][j];
37                              a++;
38                      }
39              }
40
41              b = 0;
42              flag = 0;
43
44              for(p = 0; p < a; p++)
45              {
46                      fprintf(aptr, "%c ", store[p]);
47                      store2[b] = store[p];
48                      b++;
49
50                      while(store[p] == store[p+1])
51                      {
52                              p++;
53                      }
54
55              }
56
57              fprintf(aptr, "\n");
58
59              for(f = 0; f < b; f++)
60              {
61                      if(store2[f] == 'P')
62                      {
63                              store2[f] = 'I';
64                      }
65              }
66
67              count = 0;
68
69              for(q = 0; q < b; q++)
70              {
71                      if(store2[q] == 'I')
72                      {
73                              count++;
74                      }
75              }
```

```
76
77                  if(count > 1)
78                  {
79                          fprintf(bptr, "I ");
80                          for(p = 0; p < b; p++)
81                          {
82                                  if(store2[p] != 'I')
83                                  {
84                                          fprintf(bptr, "%c ",
                                            ↪ store2[p]);
85                                  }
86                          }
87                  }
88                  else
89                  {
90                          for(p = 0; p < b; p++)
91                          {
92                                  fprintf(bptr, "%c ", store2[p]);
93                          }
94                  }
95                  fprintf(optr, "\n");
96                  fprintf(bptr, "\n");
97          }
98          return 0;
99  }
```

With thanks to Xavier Blake.

A.4 Parse tree

This parse tree was created based on the table of combinations of part-of-speech tags for the twelve grammatical tenses of declarative sentences. The parse tree is rather large and so it is necessary to split the image into two figures.

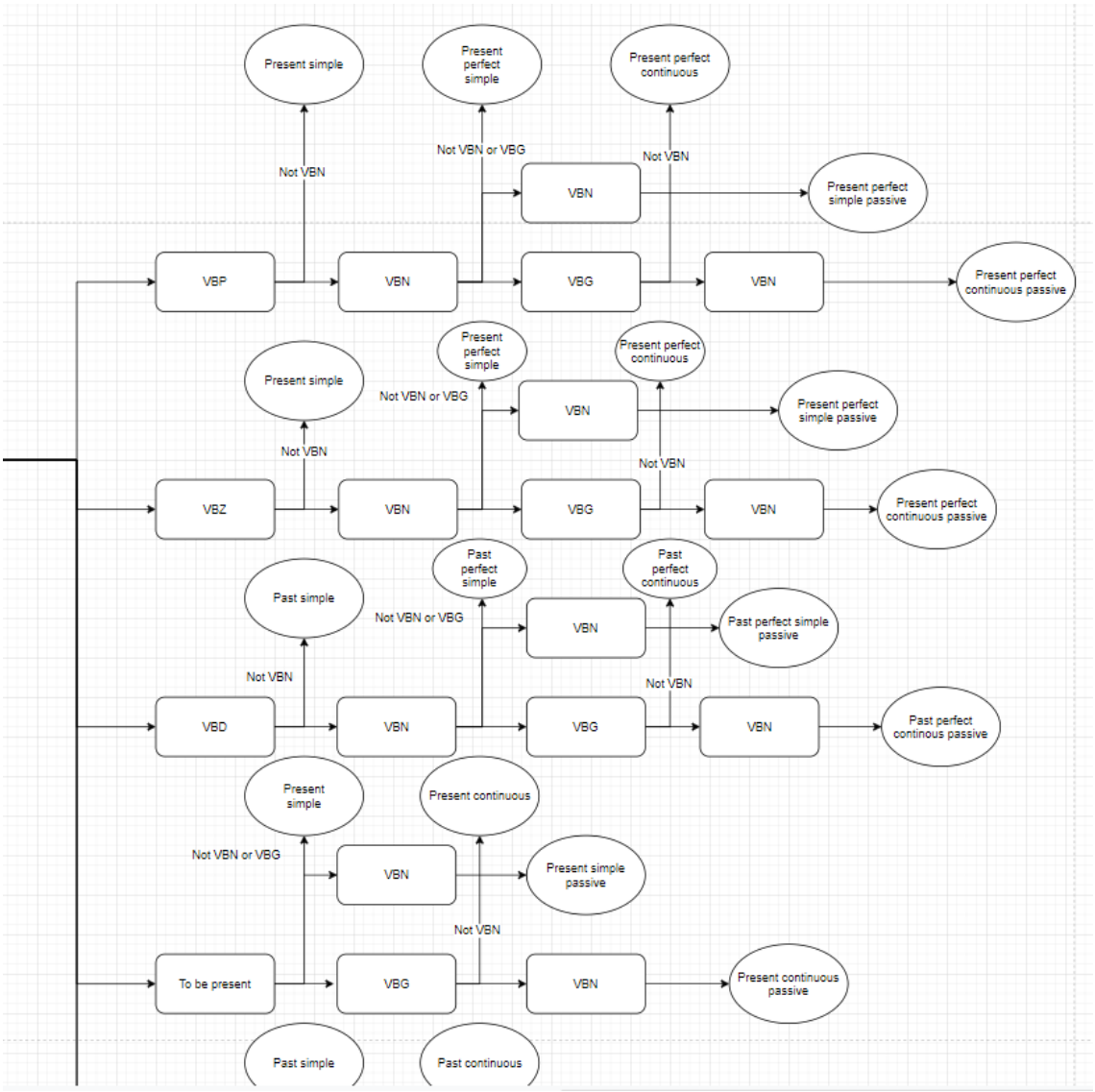


FIGURE A.1: Top half of parse tree diagram

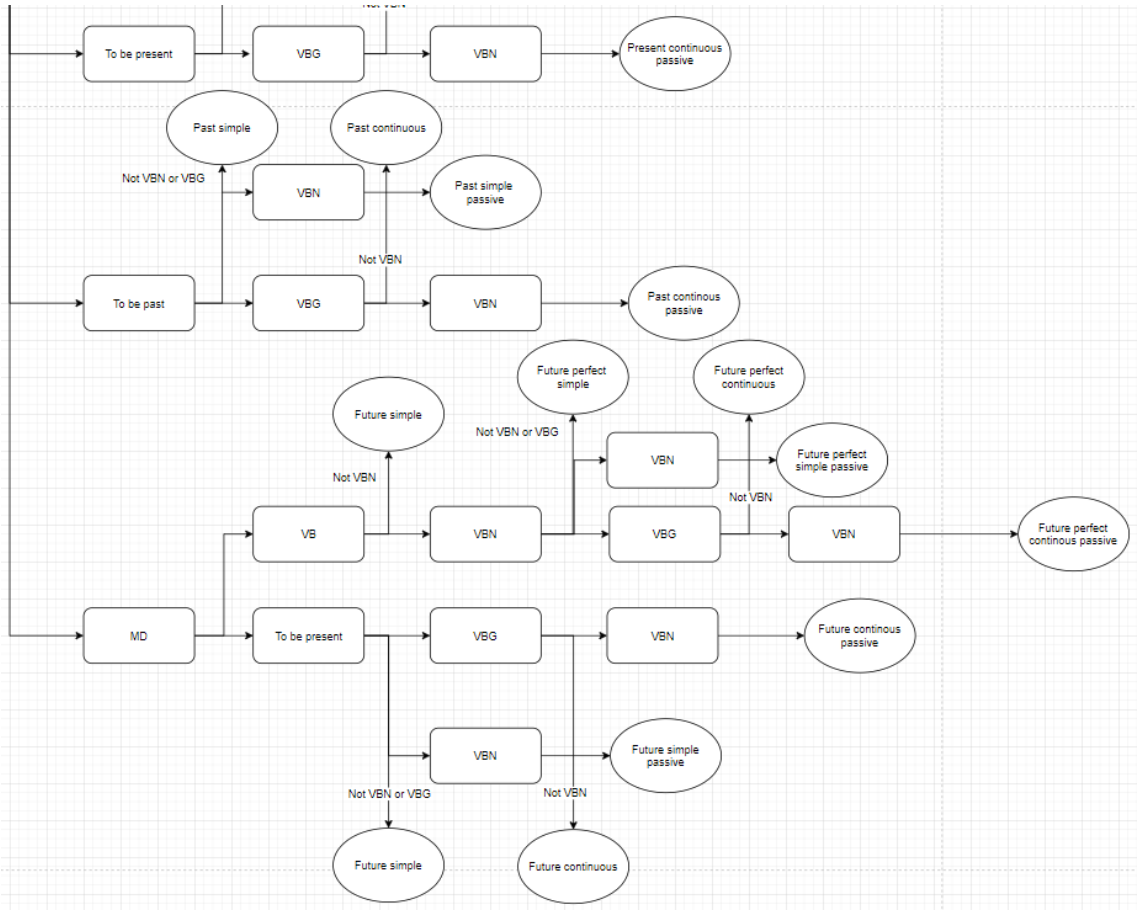


FIGURE A.2: Bottom half of parse tree diagram

A.5 R script for adjaceny pairs and heatmaps

```
createUpdatePairFreq <- function(perm_list, pair_freq) {
  for(i in 1:(length(perm_list)-1)){
    pair <- paste0(perm_list[i], perm_list[i + 1])
    if (is.null(pair_freq[[pair]])) {
      pair_freq[pair] <- 1
    }
    else {
      pair_freq[[pair]] <- pair_freq[[pair]] + 1
    }
  }
  return(pair_freq)
}
createMovePermLists <- function(perm_df) {
  move_perm_lists <-summarize(
    group_by(perm_df[,c("TextFile", "Permutation")], TextFile),
     ↪  FeatureSequence = list(Permutation)
  )
  return(move_perm_lists)
}
createDisciplinePairFreqsList <- function (move_perm_lists_df,
 ↪  disciplines_list) {
  discipline_pair_freqs <- list()
  for(discipline_name in disciplines_list) {
    single_discipline_df <- move_perm_lists_df[grep(discipline_name,
     ↪  move_perm_lists_df$TextFile), "FeatureSequence"]
    PairFreq <- list()
    for (abstract in single_discipline_df$FeatureSequence) {
      PairFreq <- createUpdatePairFreq(abstract, PairFreq)
    }
    discipline_pair_freqs[[discipline_name]] <- PairFreq
  }
  return(discipline_pair_freqs)
}
createDisciplinePairFreqsMat <- function (freqs_list,
 ↪  disciplines_list) {
  disciplines_list <- sort(disciplines_list)
  adjacent_pairs <- c(
    "IP", "IM", "IR", "ID", "PI", "PM", "PR", "PD", "MI", "MP",
    "MR", "MD", "RI", "RP", "RM", "RD", "DI", "DP", "DM", "DR"
  )
```

```
37    freq_mat <- matrix(
38      nrow = length(adjacent_pairs), ncol = length(disciplines_list),
        ↪  data = 0,
39      dimnames = list(adjacent_pairs, disciplines_list)
40    )
41    for (discipline_name in disciplines_list) {
42      for (pair in adjacent_pairs) {
43        if (!is.null(freqs_list[[discipline_name]][[pair]])) {
44          freq_mat[pair, discipline_name] <-
            ↪  freqs_list[[discipline_name]][[pair]]
45        }
46      }
47    }
48    return(freq_mat)
49  }
50
51  createMoveTransMats <- function (freq_mat, disciplines_list) {
52    disciplines_list <- sort(disciplines_list)
53    discipline_trans_mats <- list()
54    for (discipline_name in disciplines_list) {
55      disc_counts <- freq_mat[, discipline_name]
56      moves_list <- c("I", "P", "M", "R", "D")
57      trans_mat <- matrix(
58        nrow = length(moves_list), ncol = length(moves_list), data = 0,
59        dimnames = list(moves_list, moves_list)
60      )
61      for (move_pair in names(disc_counts)) {
62        move_pair_list <- unlist(strsplit(move_pair, ""))
63        trans_mat[move_pair_list[1], move_pair_list[2]] <-
          ↪  disc_counts[move_pair]
64      }
65      discipline_trans_mats[[discipline_name]] <- trans_mat
66    }
67    return(discipline_trans_mats)
68  }
```

A.6 R script for actual vs potential permutations

```
createPotRealPermText <- function (move_perm_list_ds, colour,
↪  ln_break_count) {
  counter <- 1
  pot_real_perm_text <- ""
  for (perm in permn(c("I", "P", "M", "R", "D"))) {
    perm_text <- paste0(unlist(perm), collapse='')
    found_perm <- FALSE
    for (seq in move_perm_list_ds$FeatureSequence) {
      if (paste0(unlist(seq), collapse='') == perm_text) {
        found_perm <- TRUE
        break
      }
    }
    if (found_perm) {
      pot_real_perm_text <- paste0(pot_real_perm_text, "\\color{",
      ↪  colour, "}{", perm_text, "}")
    }
    else {
      pot_real_perm_text <- paste0(pot_real_perm_text, perm_text)
    }
    if (counter == ln_break_count) {
      pot_real_perm_text <- paste0(pot_real_perm_text, " \\\\\n")
      counter <- 1
    }
    else {
      pot_real_perm_text <- paste0(pot_real_perm_text, " & ")
      counter <- counter + 1
    }
  }
  return(pot_real_perm_text)
}
```

A.7 Box plot script

```
% Calculate means and medians
disc_avg_med = pd.DataFrame(columns=["Mean", "Median"])
disc_avg_med.Mean =
 ↪  df.groupby('Discipline').mean('SentenceLength')['SentenceLength'].round(0).astype(int
disc_avg_med.Median =
 ↪  df.groupby('Discipline').median('SentenceLength')['SentenceLength'].astype(int)
disc_avg_med = disc_avg_med.transpose()

% Create LaTeX table for first five disciplines
first_disc_avg_med = disc_avg_med.loc[:,"BOT":"IT"]
print(first_disc_avg_med.to_latex())
first_disc_avg_med

% Create LaTeX table for second five disciplines
second_disc_avg_med = disc_avg_med.loc[:,"KDE":"WC"]
print(second_disc_avg_med.to_latex())
second_disc_avg_med

% Create box plot without outliers
show_outliers = False
axes = df.boxplot('SentenceLength', rot=45, by='Discipline',
 ↪  showfliers=show_outliers)
boxplot_title = 'Sentence Length'
plt.title(boxplot_title)
plt.suptitle('')
plt.savefig("sentence_lengths_box.png")
plt.show()
```

A.8 Clustering keyness and tense features

```
import numpy as np
from sklearn.preprocessing import normalize
from sklearn.cluster import k_means
from sklearn.manifold import MDS

def get_keyness_features(categories_list, keyness_dict):
    keyness_disc_dict = {}
    for disc in categories_list:
        keyness_disc_dict[disc] = {
            tok_keyness['item'][0]: tok_keyness['keyness'] for
            ↪ tok_keyness in keyness_dict[disc]['all']
        }

    disc_tok_intersection =
    ↪ set.intersection(*[set(keyness_disc.keys()) for keyness_disc
    ↪ in keyness_disc_dict.values()])

    keyness_features = np.zeros([len(categories_list),
    ↪ len(disc_tok_intersection)])
    for disc_idx, disc in enumerate(categories_list):
        for tok_idx, tok in enumerate(disc_tok_intersection):
            keyness_features[disc_idx, tok_idx] =
            ↪ keyness_disc_dict[disc][tok]
    return normalize(keyness_features)

def plot_clusters_scatter(features, four_cluster_groups, disc_names,
↪ title, filename=None):
    colors = ['red', 'green', 'blue', 'yellow']

    embedding = MDS(n_components=2, random_state=0)
    mds_embedding = embedding.fit_transform(features)

    fig, ax = plt.subplots()
    ax.grid()
    ax.set_title(title)
    ax.set_xlabel("Dim. 1")
    ax.set_ylabel("Dim. 2")
    x_offset = 0.08 * (np.max(mds_embedding[:, 0]) -
    ↪ np.min(mds_embedding[:, 0]))
```

```
33      y_offset = 0.08 * (np.max(mds_embedding[:, 1]) -
        ↪ np.min(mds_embedding[:, 1]))
34      for idx, disc in enumerate(disc_names):
35          ax.scatter(mds_embedding[idx, 0], mds_embedding[idx, 1],
            ↪ color=colors[four_cluster_groups[idx]])
36          ax.annotate(
37              disc, (mds_embedding[idx, 0], mds_embedding[idx, 1]),
38              xytext=(mds_embedding[idx, 0]+x_offset, mds_embedding[idx,
                ↪ 1]-y_offset),
39              arrowprops = dict(arrowstyle="-",
                ↪ connectionstyle="angle3,angleA=0,angleB=-90")
40          )
41      if filename is not None:
42          fig.savefig(filename)
43      fig.show()
44
45  disc_keyness_features = get_keyness_features(disciplines,
    ↪ keyness_disciplines_data)
46  print("Number of Keywords: {}".format(disc_keyness_features.shape[1]))
47  k_groups = k_means(disc_keyness_features, 4, random_state=0)[1]
48  print(k_groups)
49  fname = "disc_keyness_clusters.png"
50  scatterplot_title = 'Disciplines in Keyness Feature space'
51  plot_clusters_scatter(disc_keyness_features, k_groups, disciplines,
    ↪ scatterplot_title, filename=fname)
52
53  tense_features = normalize(np.transpose(tense_disc_counts.to_numpy()))
54  t_groups = k_means(tense_features, 4, random_state=0)[1]
55  print(t_groups)
56  fname = "disc_tense_clusters.png"
57  scatterplot_title = 'Disciplines in Tense Feature space'
58  plot_clusters_scatter(tense_features, t_groups, disciplines,
    ↪ scatterplot_title, filename=fname)
59
60  tense_keyness_features = normalize(np.concatenate((tense_features,
    ↪ disc_keyness_features), axis=1))
61  tk_groups = k_means(tense_keyness_features, 4, random_state=0)[1]
62  print(tk_groups)
63  fname = "disc_tense_keyness_clusters.png"
64  scatterplot_title = 'Disciplines in Tense & Keyness Feature space'
65  plot_clusters_scatter(tense_keyness_features, tk_groups, disciplines,
    ↪ scatterplot_title, filename=fname)
```

```
66
67 move_keyness_features = get_keyness_features(moves,
   ↪ keyness_moves_data)
68 print("Number of Keywords: {}".format(move_keyness_features.shape[1]))
69 k_groups = k_means(move_keyness_features, 4, random_state=0)[1]
70 print(k_groups)
71 fname = "move_keyness_clusters.png"
72 scatterplot_title = 'Moves in Keyness Feature space'
73 plot_clusters_scatter(move_keyness_features, k_groups, moves,
   ↪ scatterplot_title, filename=fname)
74
75 tense_features = normalize(np.transpose(tense_move_counts.to_numpy()))
76 t_groups = k_means(tense_features, 4, random_state=0)[1]
77 print(t_groups)
78 fname = "move_tense_clusters.png"
79 scatterplot_title = 'Moves in Tense Feature space'
80 plot_clusters_scatter(tense_features, t_groups, moves,
   ↪ scatterplot_title, filename=fname)
81
82 tense_keyness_features = normalize(np.concatenate((tense_features,
   ↪ move_keyness_features), axis=1))
83 tk_groups = k_means(tense_keyness_features, 4, random_state=0)[1]
84 print(tk_groups)
85 fname = "move_tense_keyness_clusters.png"
86 scatterplot_title = 'Moves in Tense & Keyness Feature space'
87 plot_clusters_scatter(tense_keyness_features, tk_groups, moves,
   ↪ scatterplot_title, filename=fname)
```

A.9 Keyness script

```
git clone git@github.com:JasperD-UGent/keyness-calculator.git

os.chdir('keyness-calculator')

from utils import init_keyness_calculator

def get_keyness_pos(treebank_tag):
    if treebank_tag.startswith('J'):
        return "ADJ"
    elif treebank_tag.startswith('V'):
        return "VERB"
    elif treebank_tag.startswith('N'):
        return "NOUN"
    elif treebank_tag.startswith('R'):
        return "ADV"
    else:
        return

def get_wordnet_pos(treebank_tag):
    if treebank_tag.startswith('J'):
        return wordnet.ADJ
    elif treebank_tag.startswith('V'):
        return wordnet.VERB
    elif treebank_tag.startswith('N'):
        return wordnet.NOUN
    elif treebank_tag.startswith('R'):
        return wordnet.ADV
    else:
        return ''

%backslashes added before dollar sign and percent sign in line 43 to avoid problem in LaTeX
def prep_keyness_text(txt, filter_punctuation=True):
    tokens = WordPunctTokenizer().tokenize(txt)
    tokens = pos_tag(tokens)
    tokens = manual_pos_correction(tokens)
    if filter_punctuation:
        tok_filter = "°~!@#\$\%^&*()_+{}|:\"<>?`-=[]\;',./"
        tokens = [tok for tok in tokens if tok[0] and tok[0][0] not in tok_filter]
```

```
39      lemmas = []
40      for tok in tokens:
41          wn_pos = get_wordnet_pos(tok[1])
42          if wn_pos:
43              lemmas.append(WordNetLemmatizer().lemmatize(tok[0],
                ↪  wn_pos))
44          else:
45              lemmas.append(WordNetLemmatizer().lemmatize(tok[0]))
46      return [(tok[0].lower(), get_keyness_pos(tok[1]), lemma.lower())
        ↪  for tok, lemma in zip(tokens, lemmas)]
47
48  def generate_keyness_data(disc_names, disc_text_df, metric="Ratio",
    ↪  pos=("NOUN", "ADJ", "VERB", "ADV"), per_move=False):
49      keyness_dictionary = dict()
50      input_rc = (
51          "RC_all_disciplines",
52          {
53              "RC_all_disc_corpus": [
54                  prep_keyness_text(text) for text
55                  in disc_text_df.groupby("AbstractName").Text.apply("
                    ↪  ".join)
56              ]
57          }
58      )
59      for discipline in disc_names:
60          if per_move:
61              filt = 'Move'
62              print("Generating Keyness statistics for {}
                ↪  move.".format(discipline))
63          else:
64              filt = 'Discipline'
65              print("Generating Keyness statistics for {}
                ↪  discipline.".format(discipline))
66          input_sc = (
67              "SC_{}".format(discipline),
68              {
69                  "SC_{}_corpus".format(discipline): [
70                      prep_keyness_text(text) for text
71                      in disc_text_df[disc_text_df[filt] ==
                        ↪  discipline].groupby("AbstractName").Text.apply("
                        ↪  ".join)
72                  ]
```

```
            }
        )
        keyness_dictionary[discipline] = init_keyness_calculator(
            input_sc, input_rc, keyness_metric=metric, desired_pos=pos
        )
    return keyness_dictionary

disciplines = list(discipline_names.values())
moves = ['introduction', 'purpose', 'method', 'result', 'discussion']

disc_mv_sent_df = df[["Discipline", "AbstractName", "Move", "Text"]]
del df

keyness_disciplines_data = generate_keyness_data(disciplines,
↪  disc_mv_sent_df, pos=("ADV",))
keyness_moves_data = generate_keyness_data(moves, disc_mv_sent_df,
↪  pos=("ADV",), per_move=True)
```

A.10 Sentence length script

```
disc_avg_med = pd.DataFrame(columns=["Mean", "Median"])
disc_avg_med.Mean =
↪  df.groupby('Discipline').mean('SentenceLength')['SentenceLength'].round(0).astype
disc_avg_med.Median =
↪  df.groupby('Discipline').median('SentenceLength')['SentenceLength'].astype(int)
disc_avg_med = disc_avg_med.transpose()

first_disc_avg_med = disc_avg_med.loc[:,"BOT":"IT"]
print(first_disc_avg_med.to_latex())
first_disc_avg_med

second_disc_avg_med = disc_avg_med.loc[:,"KDE":"WC"]
print(second_disc_avg_med.to_latex())
second_disc_avg_med

show_outliers = False
axes = df.boxplot('SentenceLength', rot=45, by='Discipline',
↪  showfliers=show_outliers)
boxplot_title = 'Sentence Length'
plt.title(boxplot_title)
plt.suptitle('')
plt.savefig("sentence_lengths_box.png")
plt.show()
```

A.11 Tense identification script

```
import re

import nltk
from nltk import WordPunctTokenizer
from anytree.search import findall_by_attr
from anytree import Node


class TreeModel:
    root = Node("root")
    # root children
    vbp = Node("vbp", parent=root, tense="pressimp")
    vbz = Node("vbz", parent=root, tense="pressimp")
    vbd = Node("vbd", parent=root, tense="pastsimp")
    tobePres = Node("tobepres", parent=root, tense="pressimp")
    tobePast = Node("tobepast", parent=root, tense="pastsimp")
    md = Node("md", parent=root, tense=None)

    # vbp children
    vbp_vb = Node("vb", parent=vbp, tense="pressimp")
    vbp_vbn = Node("vbn", parent=vbp, tense="presperfsimp")
    vbp_vbn_vbn = Node("vbn", parent=vbp_vbn,
    ↪ tense="presperfsimppass")
    vbp_vbn_vbg = Node("vbg", parent=vbp_vbn, tense="presperfcont")
    vbp_vbn_vbg_vbn = Node("vbn", parent=vbp_vbn_vbg,
    ↪ tense="presperfcontpass")

    # vbz children
    vbz_vb = Node("vb", parent=vbz, tense="pressimp")
    vbz_vbn = Node("vbn", parent=vbz, tense="presperfsimp")
    vbz_vbn_vbn = Node("vbn", parent=vbz_vbn,
    ↪ tense="presperfsimppass")
    vbz_vbn_vbg = Node("vbg", parent=vbz_vbn, tense="presperfcont")
    vbz_vbn_vbg_vbn = Node("vbn", parent=vbz_vbn_vbg,
    ↪ tense="presperfcontpass")

    # vbd children
    vbd_vb = Node("vb", parent=vbd, tense="pastsimp")
    vbd_vbn = Node("vbn", parent=vbd, tense="pastperfsimp")
```

```
36      vbd_vbn_vbn = Node("vbn", parent=vbd_vbn,
        ↪  tense="pastperfsimppass")
37      vbd_vbn_vbg = Node("vbg", parent=vbd_vbn, tense="pastperfcont")
38      vbd_vbn_vbg_vbn = Node("vbn", parent=vbd_vbn_vbg,
        ↪  tense="pastperfcontpass")
39
40      # tobePres children
41      tobePres_vbn = Node("vbn", parent=tobePres, tense="pressimppass")
42      tobePres_vbg = Node("vbg", parent=tobePres, tense="prescont")
43      tobePres_vbg_vbn = Node("vbn", parent=tobePres_vbg,
        ↪  tense="prescontpass")
44
45      # tobePast children
46      tobePast_vbn = Node("vbn", parent=tobePast, tense="pastsimppass")
47      tobePast_vbg = Node("vbg", parent=tobePast, tense="pastcont")
48      tobePast_vbg_vbn = Node("vbn", parent=tobePast_vbg,
        ↪  tense="pastcontpass")
49
50      # md children
51      md_vb = Node("vb", parent=md, tense="futusimp")
52      md_vb_vbn = Node("vbn", parent=md_vb, tense="futrperfsimp")
53      md_vb_vbn_vbn = Node("vbn", parent=md_vb_vbn,
        ↪  tense="futuperfsimppass")
54      md_vb_vbn_vbg = Node("vbg", parent=md_vb_vbn,
        ↪  tense="futuperfcont")
55      md_vb_vbn_vbg_vbn = Node("vbn", parent=md_vb_vbn_vbg,
        ↪  tense="futuperfcontpass")
56
57      md_tobePres = Node("tobepres", parent=md, tense="futusimp")
58      md_tobePres_vbg = Node("vbg", parent=md_tobePres,
        ↪  tense="futucont")
59      md_tobePres_vbg_vbn = Node("vbn", parent=md_tobePres_vbg,
        ↪  tense="futucontpass")
60      md_tobePres_vbn = Node("vbn", parent=md_tobePres,
        ↪  tense="futusimppass")
61
62
63  def check_to_be(token):
64      tobe_pres = ['is', 'are', 'am']
65      tobe_past = ['was', 'were']
66      temp = list(token)
67      if temp[0].lower() in tobe_pres:
```

```
68          temp[1] = "tobepres"
69      if temp[0].lower() in tobe_past:
70          temp[1] = "tobepast"
71      return tuple(temp)
72
73
74  def get_verbs(tokens, tokens_idx, verbose=False):
75      verbs_list = ['VBP', 'VBN', 'VBZ', 'VBD', 'VBG', 'MD', 'VB']
76
77      def equal_vb_pos(toks, idx):
78          # previous verb pos tag must equal the next one for elision
            ↪ condition to pass
79          if idx < 1:
80              return False
81          last_verb_pos = toks[idx - 1][1]
82          next_verb_pos = ""
83          for tok in toks[idx + 1:]:
84              if tok[1] in verbs_list:
85                  next_verb_pos = tok[1]
86                  break
87          return last_verb_pos == next_verb_pos
88
89      verb_groups = []
90      verbs = []
91      i = 0
92      while i < len(tokens):
93          token = tokens[i]
94          if verbose:
95              print("token in get_verbs: ", token)
96          if token[1] in verbs_list:
97              token += ([tokens_idx[i][0], tokens_idx[i][1]],)
98              if verbose:
99                  print("in if, token: ", token)
100             verbs.append(token)
101         elif (token[1] == ',' or token[1] == 'CC') and
            ↪ equal_vb_pos(tokens, i):  # elisions - conjunctions
            ↪ between verbs
102             pass
103         elif re.match("RB*", token[1]):  # adjective between verbs
104             pass
105         elif re.match("NN*", token[1]) or re.match("PRP*", token[1]):
            ↪ # noun or personal pronoun between verbs
```

```
106             pass
107         else:
108             if len(verbs) != 0:
109                 verb_groups.append(verbs)
110             verbs = []
111         i += 1
112     if len(verbs) != 0:
113         verb_groups.append(verbs)
114     for verb_group in verb_groups:
115         for i in range(len(verb_group)):
116             verb_group[i] = check_to_be(verb_group[i])
117     return verb_groups
118 
119 
120 def get_tense(tokens, verbose=False):
121     starting_node = TreeModel.root
122     mark = 0
123     for i in range(len(tokens)):
124         token = tokens[i]
125         if verbose:
126             print("tokens: ", tokens)
127             print("token: ", token)
128         found = findall_by_attr(starting_node, token[1].lower(),
            ↪ maxlevel=2)
129         if len(found) != 0:
130             starting_node = found[0]
131             if i == len(tokens) - 1:
132                 for j in range(mark, i + 1):
133                     tense = starting_node.tense
134                     depth = starting_node.depth
135                     tokens[j] += (tense, depth,)  # ',' to make tuple
136         else:
137             if starting_node != TreeModel.root:
138                 for j in range(mark, i):
139                     tense = starting_node.tense
140                     depth = starting_node.depth
141                     tokens[j] += (tense, depth,)  # ',' to make tuple
142                 starting_node = TreeModel.root
143                 i -= 1
144             mark = i
145     return tokens
146 
```

```
147
148 def manual_pos_correction(tokens):
149     corrections = {
150         "Did": ('Did', 'VBD'),
151         "Are": ('Are', 'VBP'),
152     }
153     for idx, token in enumerate(tokens):
154         if token[0] in corrections:
155             tokens[idx] = corrections[token[0]]
156
157     return tokens
158
159
160 def get_tense_verb_groups(text, verbose=False):
161     tokens = WordPunctTokenizer().tokenize(text)
162     if verbose:
163         print("**************", len(tokens))
164     tokens_idx = list(WordPunctTokenizer().span_tokenize(text))
165     tokens = nltk.pos_tag(tokens)
166     tokens = manual_pos_correction(tokens)
167     if verbose:
168         print("================", len(tokens), len(tokens_idx))
169         print("tokens: ", tokens)
170     verb_groups = get_verbs(tokens, tokens_idx)
171     if verbose:
172         print("verb_groups: ", verb_groups)
173     tense_verb_groups = []
174     for verb_group in verb_groups:
175         tense_verb_group = get_tense(verb_group, verbose=verbose)
176         tense_verb_groups.append(tense_verb_group)
177     return tense_verb_groups
178
179
180 def print_parsed_text(text):
181     tense_vb_groups = get_tense_verb_groups(text, verbose=False)
182     # initialize counts for all possible labels
183     all_tenses = {
184         'futrperfsimp', 'futucont', 'futucontpass', 'futuperfcont',
185         'futuperfcontpass', 'futuperfsimppass', 'futusimp',
            ↪ 'futusimppass',
186         'pastcont', 'pastcontpass', 'pastperfcont',
            ↪ 'pastperfcontpass',
```

```
187         'pastperfsimp', 'pastperfsimppass', 'pastsimp',
            ↪ 'pastsimppass',
188         'prescont', 'prescontpass', 'presperfcont',
            ↪ 'presperfcontpass',
189         'presperfsimp', 'presperfsimppass', 'pressimp', 'pressimppass'
190     }
191     labels_elisions_counts = {tense: {"labels": 0, "elisions": 0} for
        ↪ tense in all_tenses}
192
193     out_txt = text
194     span_correction = 0
195     for vb_group in tense_vb_groups:
196         tree_depth = 0
197         tense_vb_count = 0
198         last_token_span = []
199         vb_tense = ""
200         for verb in vb_group:
201             if len(verb) == 5:  # verb with a successfully classified
                ↪ tense
202                 last_token_span, vb_tense, tree_depth = verb[2:]
203                 tense_vb_count += 1
204         if tense_vb_count != 0:
205             labels_elisions_counts[vb_tense]["labels"] += 1
206             labels_elisions_counts[vb_tense]["elisions"] +=
                ↪ tense_vb_count - tree_depth
207             insertion_index = last_token_span[1] + span_correction
208             insertion = '<{}>'.format(vb_tense)
209             out_txt = out_txt[:insertion_index] + insertion +
                ↪ out_txt[insertion_index:]
210             span_correction += len('<{}>'.format(vb_tense))
211
212     print(format_text_output(all_tenses, labels_elisions_counts,
        ↪ out_txt))
213
214
215 def format_text_output(all_tenses, labels_elisions_nbs, output_text):
216     print("\n" + "-" * 79 + "\n" + output_text)
217     label_margin = max([len(tense) for tense in all_tenses])
218     counts_txt = "-" * 79
219     for tense, counts in sorted(labels_elisions_nbs.items()):
220         counts_txt += "\n" + tense + ": " + " " * (label_margin -
            ↪ len(tense))
```

```
221         for count_name, count in counts.items():
222             counts_txt += "{} {}; ".format(count, count_name)
223     return counts_txt
224 
225 
226 if __name__ == "__main__":
227     import argparse
228 
229     parser = argparse.ArgumentParser(
230         description='Tool to classify tenses and count the number of
            ↪ tense labels & elisions.'
231     )
232     parser.add_argument('text_path', help='A path to the text file to
        ↪ be processed.')
233     args = parser.parse_args()
234 
235     with open(args.text_path, 'r') as f:
236         txt = f.read()
237 
238     #### Some sample test items ####
239     # txt = "The item is packed, is checked and is delivered."
240     # txt = "The item is packed, checked and delivered."
241     # txt = "It had been working and functioning properly."
242     # txt = "Did anything happen? Are you okay? When do you want to
        ↪ start? Is he OK?"
243     # txt = "He does not know!"
244 
245     print_parsed_text(txt)
```

A.12 Tense count script

```
import os
from xml.dom.minidom import parse as parse_xml

import pandas as pd
import matplotlib.pyplot as plt
from nltk import WordPunctTokenizer, pos_tag
from nltk.stem import WordNetLemmatizer
from nltk.corpus import wordnet

from black.black_task import get_tense_verb_groups,
    manual_pos_correction

dataset_path = "..."

def extract_segments(folder_path, file_name):
    file_path = os.path.join(folder_path, file_name)
    raw_segments =
        parse_xml(file_path).getElementsByTagName("segment")
    segments_list = []
    for segment in raw_segments:
        sentence = segment.firstChild.data
        features = segment.attributes["features"].firstChild.data
        idx = segment.attributes["id"].firstChild.data
        move_submove = features.split(";")[1:]
        if len(move_submove) == 1:
            move, submove = move_submove[0], None
        else:
            move, submove = move_submove
        segments_list.append((idx, move, submove, sentence))
    return segments_list

discipline_names = {
    "Bot_": "BOT",
    "Tr_on_EvolutionaryComp_": "EC",
    "IE_": "IND",
    "Tr_on_ImageProcess_": "IP",
    "Tr_on_In_Th_": "IT",
    "Tr_on_Know&DataEng_": "KDE",
    "Ling_": "LING",
    "Mat_":"MAT",
```

```
    "Med_": "MED",
    "Tr_on_WirelessComm_": "WC",
}
def get_discipline(abstract_name):
    for discipline_file_name, discipline_name in discipline_names.items():
        if discipline_file_name in abstract_name:
            return discipline_name
    return None

def get_sent_tense(sent_txt, verbose=True):
    tense_verb_groups = get_tense_verb_groups(sent_txt, verbose=verbose)
    final_verbs = []
    for verb_group in tense_verb_groups:
        valid_verbs = [verb for verb in verb_group if len(verb) == 5] # verbs with a classified tense only
        if valid_verbs:
            tree_depth = valid_verbs[0][4]
            nb_aux = tree_depth - 1
            nb_main = len(valid_verbs) - nb_aux
            # split verbs into auxiliary and main, remove the tree depth information
            # resulting format: ('verb', 'POS', ['L span', 'R span'], 'tense')
            aux_verbs = [verb[:4] for verb in valid_verbs[:nb_aux]]
            main_verbs = [verb[:4] for verb in valid_verbs[-nb_main:]]
            final_verbs.extend(aux_verbs)
            final_verbs.extend(main_verbs)
    first_vb_tense = final_verbs[0][3] if final_verbs else None
    return first_vb_tense

tense_dictionary = {
    'futuperfsimp': 'Future Perfect Simple',
    'futucont': 'Future Continuous',
    'futuperfcont': 'Future Perfect Continuous',
    'futusimp': 'Future Simple',
    'pastcont': 'Past Continuous',
    'pastperfcont': 'Past Perfect Continuous',
    'pastperfsimp': 'Past Perfect Simple',
    'pastsimp': 'Past Simple',
    'pastsimppass': 'Past Simple',
```

```
    'prescont': 'Present Continuous',
    'prescontpass': 'Present Continuous',
    'presperfcont': 'Present Perfect Continuous',
    'presperfsimp': 'Present Perfect Simple',
    'pressimp': 'Present Simple',
    'pressimppass': 'Present Simple',
    None: None
}

def dataset_segment_iterator(ds_path, filter_punctuation=True):
    for f_name in os.listdir(ds_path):
        abstract_name = f_name.split('.')[0]
        segments = extract_segments(ds_path, f_name)
        discipline = get_discipline(abstract_name)
        for segment in segments:
            sent_text = segment[3]
            tok_filter = "°~!@#$%^&*()_+{}|:\"<>?`-=[]\;',./"
            if filter_punctuation:
                sent_len = len([tok for tok in
                ↪ WordPunctTokenizer().tokenize(sent_text) if tok
                ↪ not in tok_filter])
            else:
                sent_len =
                ↪ len(WordPunctTokenizer().tokenize(sent_text))
            sent_tense = tense_dictionary[get_sent_tense(sent_text,
            ↪ verbose=False)]
            yield {
                'AbstractName': abstract_name,
                'Discipline': discipline,
                'SentenceID': segment[0],
                'Move': segment[1],
                'Submove': segment[2],
                'Tense': sent_tense,
                'SentenceLength': sent_len,
                'Text': sent_text,
            }

df = pd.DataFrame(
    [seg for seg in dataset_segment_iterator(dataset_path,
    ↪ filter_punctuation=True)]

tense_disc_counts = df.groupby("Discipline").Tense.value_counts()
```

```
tense_disc_counts =
    ↪ tense_disc_counts.to_frame().unstack(fill_value=0).transpose()
print(tense_disc_counts.to_latex())
tense_disc_counts
)
```

A.13 Multidimensional scaling and Hierarchical clustering script

This script was created to compare the manual grouping and automated clustering.

```
# comparison of disciplines based on linearity, cyclicity and variety
disciplines <- c("BOT", "EC", "IND", "IP", "IT", "KDE", "LING", "MAT", "MED", "WC")
# disc_groups <- c("2" , "4" , "4" , "3" , "4", "3" , "2" , "1" , "1" , "4" )  # using k-means instead
linear_cycl_df <- linear_cycl_Five
discipline_clv_df <- data.frame(
  Discipline=character(), Linearity=integer(), Cyclicity=integer(), Variety=integer(), stringsAsFactors=FALSE
)
for(discipline_name in sort(disciplines)) {
  one_disc_df <- linear_cycl_df[grep(discipline_name, linear_cycl_df$TextFile), ]
  linear_cycl_sums <- colSums(one_disc_df[grep(discipline_name, one_disc_df$TextFile), c("Linearity", "Cyclicity")])
  variety <- length(unique(one_disc_df$Permutation))
  disc_row <- c(discipline_name, linear_cycl_sums[["Linearity"]], linear_cycl_sums[["Cyclicity"]], variety)
  discipline_clv_df[nrow(discipline_clv_df) + 1,] <- disc_row
}

set.seed(0)
disc_groups <- as.character(kmeans(discipline_clv_df[2:4], 4)[["cluster"]])

mds <- as_tibble(cmdscale(dist(discipline_clv_df[, 2:4])))
colnames(mds) <- c("Dim. 1", "Dim. 2")
mds["Discipline"] <- discipline_clv_df$Discipline
mds["Group"] <- disc_groups

library(ggrepel)
mds_plot <- ggplot(mds, aes(x = `Dim. 1`, y = `Dim. 2`)) + geom_point()
mds_plot + geom_text_repel(aes(label = Discipline, colour = Group), size = 3)

hierarchical_clust <- hclust(dist(discipline_clv_df[, 2:4]))
```

```
29 hierarchical_clust[["labels"]] <- discipline_clv_df$Discipline
30 plot(hierarchical_clust)
```

A.14 R functions

The following functions are used as snippets that combine together to create discrete scripts, shown in the Appendixes for the R scripts.

```
library(XML)
library(stringr)
library(dplyr)

createRawMasterDf <- function(xml_path, xml_files){

  gbdf <- data.frame()

  for(i in 1:length(xml_files)){
    #for(i in 1:3){
    indata <- xmlParse(file.path(getwd(), xml_path, xml_files[i]))
    #indata =
     ↪  xmlParse(paste0(getwd(),"/data/AllText/","Tr_on_EvolutionaryComp_27.xml",""))

    xml_data_header <- xmlToList(indata)[["header"]]
    xml_data_body <- xmlToList(indata)[["body"]]

    txtfile <- xml_data_header$textfile

    # Segments
    len_seg <- length(xml_data_body)

    seg <- character(len_seg)
    feat <- character(len_seg)

    #j =5
    for(j in 1:len_seg){
      # print("####################")
      # print(i)
      # print(j)
      # print("####################")

      # Breaking out if the element has no segment
      if(is.null(xml_data_body[j][["segment"]])){
        seg <- NA
        feat <- NA
        break()
      }
```

```
38
39        if(!("text" %in% names(xml_data_body[j][["segment"]]))){
40          seg[j] <- NA
41        }
42        else{
43          seg[j] <- xml_data_body[j][["segment"]][["text"]]
44        }
45
46        if(
47          !(".attrs" %in% names(xml_data_body[j][["segment"]])) ||
48             !("features" %in%
                ↪  names(xml_data_body[j][["segment"]][[".attrs"]]))
49        ){
50          feat[j] <- NA
51        }
52        else{
53          feat[j] <-
             ↪  xml_data_body[j][["segment"]][[".attrs"]]["features"]
54        }
55
56      }
57
58      tdf <- NULL
59      tdf <- data.frame(TextFile = rep(txtfile, len_seg),
60                        Segment = seg,
61                        Features = feat,
62                        stringsAsFactors = FALSE)
63
64
65
66      gbdf <- rbind(gbdf, tdf)
67    }
68    return(gbdf)
69  }
70
71  createMasterDf <- function(rmdf){
72    rmdf$TextFile <- as.character(lapply(rmdf$TextFile, function(x){
73      pos <- str_locate(x, fixed("/"))[1]
74
75      str <- substr(x, pos+1, nchar(x))
76      str <- gsub("Tr_on_", "", str)
77
```

```
78  }))
79
80  # Removing "rhetorical_moves" from features column
81  rmdf$Features <- as.character(lapply(rmdf$Features, function(x){
82    str <- gsub("rhetorical_moves", "", x)
83    str <- trimws(gsub(";", " ", str))
84  }))
85
86  # Adding Moves and Submoves columns
87  master <- rmdf
88  #master = data.frame(rmdf[,],Moves = rep("NULL",nrow(rmdf)),Submoves
    ↪  = rep("NULL",nrow(rmdf)))
89
90  for(i in 1:nrow(master)){
91
92    # Filenames
93    if(length(grep("EvolutionaryComp",master[i,1]))>0){
94      master[i,1] <- gsub("EvolutionaryComp", "EC", master[i, 1])
95    }
96
97    if(length(grep("WirelessComm",master[i,1]))>0){
98      master[i,1] <- gsub("WirelessComm", "WC", master[i, 1])
99    }
100
101   if(length(grep("Know&DataEng",master[i,1]))>0){
102     master[i,1] <- gsub("Know&DataEng", "KDE", master[i, 1])
103   }
104
105   if(length(grep("In_Th",master[i,1]))>0){
106     master[i,1] <- gsub("In_Th", "IT", master[i, 1])
107   }
108
109   if(length(grep("ImageProcess",master[i,1]))>0){
110     master[i,1] <- gsub("ImageProcess", "IP", master[i, 1])
111   }
112
113   if(length(grep("Bot",master[i,1]))>0){
114     master[i,1] <- gsub("Bot", "BOT", master[i, 1])
115   }
116
117   if(length(grep("IE",master[i,1]))>0){
118     master[i,1] <- gsub("IE", "IND", master[i, 1])
```

```
119       }
120
121       if(length(grep("Ling",master[i,1]))>0){
122         master[i,1] <- gsub("Ling", "LING", master[i, 1])
123       }
124
125       if(length(grep("Mat",master[i,1]))>0){
126         master[i,1] <- gsub("Mat", "MAT", master[i, 1])
127       }
128
129       if(length(grep("Med",master[i,1]))>0){
130         master[i,1] <- gsub("Med", "MED", master[i, 1])
131       }
132
133     }
134     return(master)
135
136 }
137
138 createFeatureMove <- function(ftr){
139     feature_move <- data.frame(ftr[,], Moves = rep(NA, nrow(ftr)))
140     i <- 0
141     for(i in 1:nrow(feature_move)){
142       if(length(grep("background",feature_move[i,"Features"]))>0 |
        ↪  length(grep("introduction",feature_move[i,"Features"]))>0){
143         feature_move[i,"Moves"] <- "introduction"
144       }
145
146       if(length(grep("problem",feature_move[i,"Features"]))>0 |
        ↪  length(grep("introduction",feature_move[i,"Features"]))>0){
147         feature_move[i,"Moves"] <- "introduction"
148       }
149
150       if(length(grep("gap",feature_move[i,"Features"]))>0 |
        ↪  length(grep("introduction",feature_move[i,"Features"]))>0){
151         feature_move[i,"Moves"] <- "introduction"
152       }
153
154       if(length(grep("overview",feature_move[i,"Features"]))>0 |
        ↪  length(grep("introduction",feature_move[i,"Features"]))>0){
155         feature_move[i,"Moves"] <- "introduction"
156       }
```

```
157
158     if(length(grep("purpose",feature_move[i,"Features"]))>0){
159       feature_move[i,"Moves"] <- "purpose"
160     }
161
162     if(length(grep("method",feature_move[i,"Features"]))>0){
163       feature_move[i,"Moves"] <- "method"
164     }
165
166     if(length(grep("result",feature_move[i,"Features"]))>0 |
      ↪  length(grep("methodasproduct",feature_move[i,"Features"]))>0){
167       feature_move[i,"Moves"] <- "result"
168     }
169
170     if(length(grep("result",feature_move[i,"Features"]))>0 |
      ↪  length(grep("resultasproduct",feature_move[i,"Features"]))>0){
171       feature_move[i,"Moves"] <- "result"
172     }
173
174     if(length(grep("discussion",feature_move[i,"Features"]))>0){
175       feature_move[i,"Moves"] <- "discussion"
176     }
177
178     if(length(grep("uncertain",feature_move[i,"Features"]))>0){
179       feature_move[i,"Moves"] <- "uncertain"
180     }
181   }
182
183   feature_move <- feature_move[, -2]
184
185 }
186
187 createFeatureSubmove <- function(ftr){
188   # SUb move
189   feature_submove <- data.frame(ftr[,], Submoves = rep(NA, nrow(ftr)))
190   i <- 0
191   for(i in 1:nrow(feature_submove)){
192     if(length(grep("background",feature_submove[i,"Features"]))>0){
193       feature_submove[i,"Submoves"] <- "background"
194     }
195
196     if(length(grep("problem",feature_submove[i,"Features"]))>0){
```

```
197         feature_submove[i,"Submoves"] <- "problem"
198       }
199
200       if(length(grep("gap",feature_submove[i,"Features"]))>0){
201         feature_submove[i,"Submoves"] <- "gap"
202       }
203
204       if(length(grep("overview",feature_submove[i,"Features"]))>0){
205         feature_submove[i,"Submoves"] <- "overview"
206       }
207
208       if(length(grep("purpose",feature_submove[i,"Features"]))>0){
209         feature_submove[i,"Submoves"] <- ""
210       }
211
212       if(length(grep("method",feature_submove[i,"Features"]))>0){
213         feature_submove[i,"Submoves"] <- ""
214       }
215
216
        ↪  if(length(grep("methodasproduct",feature_submove[i,"Features"]))>0){
217         feature_submove[i,"Submoves"] <- "methodasproduct"
218       }
219
220
        ↪  if(length(grep("resultasproduct",feature_submove[i,"Features"]))>0){
221         feature_submove[i,"Submoves"] <- "resultasproduct"
222       }
223
224       if(length(grep("discussion",feature_submove[i,"Features"]))>0){
225         feature_submove[i,"Submoves"] <- ""
226       }
227
228       if(length(grep("uncertain",feature_submove[i,"Features"]))>0){
229         feature_submove[i,"Submoves"] <- ""
230       }
231     }
232
233     feature_submove <- feature_submove[, -2]
234   }
235
236   createDisciplineSpecificDfs <- function(disciplines_list, master_df){
```

```
237   discipline_specific_dfs <- list()
238   for(discipline_name in disciplines_list) {
239     single_discipline_df <- masterDf[grep(discipline_name,
        ↪  master_df$TextFile), c("Segment", "Features")]
240     discipline_specific_dfs[[paste(discipline_name, "Discipline", sep
        ↪  = "_")]] <- single_discipline_df
241   }
242   return(discipline_specific_dfs)
243 }
244
245 createFeatureCounts <- function(discipline_dfs){
246   feature_counts <- list()
247   for (discipline_name in names(discipline_dfs)) {
248     feature_counts[[discipline_name]] <-
        ↪  count(discipline_dfs[[discipline_name]], Features)
249   }
250   return(feature_counts)
251 }
252
253 createRawFeatPerm <- function(master_df){
254   raw_feature_permutation_ds <- summarize(
255     group_by(master_df[,c("TextFile", "Features")], TextFile),
256     FeatureSequences = paste(Features, collapse = "; ")
257   )
258   return(raw_feature_permutation_ds)
259 }
260
261 createAbbrPerm <- function(mstPer){
262   abbr <- mstPer[,]
263
264   abbr$Moves <- as.character(lapply(abbr$Moves, function(x){
265     x <- toupper(substr(x, 1, 1))
266   }))
267
268   abbr$Submoves <- as.character(lapply(abbr$Submoves, function(x){
269     x <- tolower(substr(x, 1, 1))
270   }))
271
272   abbr$Permutation <- do.call(paste, c(abbr[c("Moves", "Submoves")],
      ↪  sep = ""))
273
274   abbr <- abbr[, -c(2, 3)]
```

```
}

createMergePerm <- function(abbr){
  curr_file <- ""
  prev_file <- ""
  curr_perm <- ""
  prev_perm <- ""

  i <- 0
  for(i in 1:nrow(abbr)){
    curr_file <- abbr[i, "TextFile"]
    curr_perm <- abbr[i, "Permutation"]

    if(i == 1){
      MergedPermutation <- data.frame(TextFile = curr_file,
       ↪  Permutation = curr_perm, stringsAsFactors = FALSE)
      prev_file <- curr_file
      prev_perm <- curr_perm
    }
    else{

      if(curr_file == prev_file & curr_perm == prev_perm){
        next()
      }
      else{
        MergedPermutation[i,"TextFile"] <- curr_file
        MergedPermutation[i,"Permutation"] <- curr_perm

        prev_file <- curr_file
        prev_perm <- curr_perm
      }

    }

  }
  MergedPermutation <-
   ↪  MergedPermutation[-which(is.na(MergedPermutation$Permutation)),]
  return(MergedPermutation)
}

createFiveMovePermutation <- function(AbreDf){
  FM_df <- AbreDf[,]
```

```
315
316   FM_df$Permutation <- as.character(lapply(AbreDf$Permutation,
      ↪  function(x){
317     x <- substr(x, 1, 1)
318   }))
319
320   return(FM_df)
321 }
322
323 createFourMovePermutation <- function(FiveMoveDf,MergeDf){
324   FiveMoveDf$Permutation <-
      ↪  as.character(lapply(FiveMoveDf$Permutation, function(x){
325     if(substr(x,1,1) == "P"){
326       x <- gsub(substr(x, 1, 1), "I", x)
327     }
328     else{
329       x <- substr(x, 1, 1)
330     }
331   }))
332
333   return(rbind(FiveMoveDf,MergeDf))
334 }
335
336 getcounts_Abbr_Move <- function(AbbrDf,FiveDf){
337   count_abbr <- count(AbbrDf, "Permutation")
338   count_five <- count(FiveDf, "Permutation")
339
340   return(list(count_abbr,count_five))
341 }
342
343 getLinearityCount <- function(mrg){
344   mrg$Permutation <- as.character(lapply(mrg$Permutation, function(x){
345     x <- substr(x, 1, 1)
346   }))
347   cf <- ""
348   pf <- ""
349   cp <- ""
350   pp <- ""
351
352   i <- 0
353   for(i in 1:nrow(mrg)){
354     cf <- mrg[i, 1]
```

```
355     cp <- mrg[i, 2]
356
357     if(cf == pf & cp == pp){
358       next()
359     }
360     else{
361       if(i==1){
362         tempdf <- data.frame(TextFile = cf, Permutation = cp,
            ↪  stringsAsFactors = FALSE)
363         pf <- cf
364         pp <- cp
365       }
366       else{
367         tempdf[i,"TextFile"] <- cf
368         tempdf[i,"Permutation"] <- cp
369
370         pf <- cf
371         pp <- cp
372       }
373     }
374   }
375   tempdf <- tempdf[-which(is.na(tempdf)==TRUE),]
376
377   files <- unique(tempdf$TextFile)
378   g <- data.frame()
379   for(l in files){
380     df <- tempdf[which(tempdf$TextFile == l),]
381     f <- l
382     p <- paste(unlist(df$Permutation), collapse = " ")
383     d <- NULL
384     d <- data.frame(TextFile = f, Permutation = p, stringsAsFactors =
        ↪  FALSE)
385     g <- rbind(g, d)
386
387   }
388
389   return(g)
390 }
391
392 checkCyclicity <- function(ln_cnt){
393
394   for(i in 1:nrow(ln_cnt)){
```

```
395
396     fl <- ln_cnt[i, 1]
397     perm <- ln_cnt[i, 2]
398     perm.vector <- as.vector(unlist(str_split(ln_cnt[i,
        ↪  "Permutation"], pattern = " ")))
399
400
401     if(length(perm.vector)-4 < 0){
402       next()
403     }
404
405     for(j in 1:length(perm.vector)){
406
407       if(j +3>length(perm.vector)){
408         break()
409       }
410       else{
411         if(perm.vector[j]== perm.vector[j+2]){
412           if(perm.vector[j+1]==perm.vector[j+3]){
413             ln_cnt[i,"Cyclicity"] <- 1
414             break()
415           }
416         }
417       }
418     }
419   }
420   ln_cnt[which(is.na(ln_cnt$Cyclicity)==TRUE),"Cyclicity"] <- 0
421   return(ln_cnt)
422 }
423
424 createUpdatePairFreq <- function(perm_list, pair_freq) {
425   for(i in 1:(length(perm_list)-1)){
426     pair <- paste0(perm_list[i], perm_list[i + 1])
427     if (is.null(pair_freq[[pair]])) {
428       pair_freq[pair] <- 1
429     }
430     else {
431       pair_freq[[pair]] <- pair_freq[[pair]] + 1
432     }
433   }
434   return(pair_freq)
435 }
```

```
436 createMovePermLists <- function(perm_df) {
437   move_perm_lists <-summarize(
438     group_by(perm_df[,c("TextFile", "Permutation")], TextFile),
      ↪  FeatureSequence = list(Permutation)
439   )
440   return(move_perm_lists)
441 }
442 createDisciplinePairFreqsList <- function (move_perm_lists_df,
    ↪  disciplines_list) {
443   discipline_pair_freqs <- list()
444   for(discipline_name in disciplines_list) {
445     single_discipline_df <- move_perm_lists_df[grep(discipline_name,
      ↪  move_perm_lists_df$TextFile), "FeatureSequence"]
446     PairFreq <- list()
447     for (abstract in single_discipline_df$FeatureSequence) {
448       PairFreq <- createUpdatePairFreq(abstract, PairFreq)
449     }
450     discipline_pair_freqs[[discipline_name]] <- PairFreq
451   }
452   return(discipline_pair_freqs)
453 }
454 createDisciplinePairFreqsMat <- function (freqs_list,
    ↪  disciplines_list) {
455   disciplines_list <- sort(disciplines_list)
456   adjacent_pairs <- c(
457     "IP", "IM", "IR", "ID", "PI", "PM", "PR", "PD", "MI", "MP",
458     "MR", "MD", "RI", "RP", "RM", "RD", "DI", "DP", "DM", "DR"
459   )
460   freq_mat <- matrix(
461     nrow = length(adjacent_pairs), ncol = length(disciplines_list),
      ↪  data = 0,
462     dimnames = list(adjacent_pairs, disciplines_list)
463   )
464   for (discipline_name in disciplines_list) {
465     for (pair in adjacent_pairs) {
466       if (!is.null(freqs_list[[discipline_name]][[pair]])) {
467         freq_mat[pair, discipline_name] <-
          ↪  freqs_list[[discipline_name]][[pair]]
468       }
469     }
470   }
471   return(freq_mat)
```

```
}

createMoveTransMats <- function (freq_mat, disciplines_list) {
  disciplines_list <- sort(disciplines_list)
  discipline_trans_mats <- list()
  for (discipline_name in disciplines_list) {
    disc_counts <- freq_mat[, discipline_name]
    moves_list <- c("I", "P", "M", "R", "D")
    trans_mat <- matrix(
      nrow = length(moves_list), ncol = length(moves_list), data = 0,
      dimnames = list(moves_list, moves_list)
    )
    for (move_pair in names(disc_counts)) {
      move_pair_list <- unlist(strsplit(move_pair, ""))
      trans_mat[move_pair_list[1], move_pair_list[2]] <-
        ↪ disc_counts[move_pair]
    }
    discipline_trans_mats[[discipline_name]] <- trans_mat
  }
  return(discipline_trans_mats)
}
```

A.15 Tailor-made R script to compare and contrast

The following script harnesses the functions shown in A.14. Each script is created from a series of functions.

```
library(XML)
library(stringr)
library(dplyr)

xml_path = "data/1000/"
xml_files = list.files(paste(getwd(),xml_path, sep='/'),pattern =
 ↪  ".xml")

source("data_funct.R")

##### Compare and contrast features #####

#  SNIPPET 1 - Read xml files into dataframe
rawMasterDf = createRawMasterDf(xml_path, xml_files)

#  SNIPPET 2 - Manipulating data (removing unnecessary columns and
 ↪  simplifying terms)
masterDf = createMasterDf(rawMasterDf)

### SNIPPET 3 - Create feature only dataframe
feature = masterDf[,-c(2)]

### SNIPPET 4 - Create feature only dataframe
feature_move = createFeatureMove(feature)
feature_submove = createFeatureSubmove(feature)
```

A.16 R script for comparing and contrasting

The following script harnesses the functions shown in A.14. Each script is created from a series of functions.

```
library(XML)
library(stringr)
library(dplyr)

xml_path = "data/1000/"
xml_files = list.files(paste(getwd(),xml_path, sep='/'),pattern =
 ↪  ".xml")

source("data_funct.R")

##### Compare and contrast sequences #####

#  SNIPPET 1 - Read xml files into dataframe
rawMasterDf = createRawMasterDf(xml_path, xml_files)

#  SNIPPET 2 - Manipulating data (removing unnecessary columns and
 ↪  simplifying terms)
masterDf = createMasterDf(rawMasterDf)


### SNIPPET 3 - Create feature only dataframe
feature = masterDf[,-c(2)]

### SNIPPET 4 - Create feature only dataframe
feature_move = createFeatureMove(feature)
feature_submove = createFeatureSubmove(feature)


## Merge feature move and submove dataframes into a single dataframe
MasterPermutation <- cbind(feature_move, Submoves =
 ↪  feature_submove$Submoves)


## SNIPPET 8 - Abbreviate feature permutations
AbbreviatedPermutation = createAbbrPerm(MasterPermutation)

## SNIPPET 9 - Merge sequential identical permutations
MergedPermutation = createMergePerm(AbbreviatedPermutation)
```

```
36
37 ## SNIPPET 10 - Omit submoves in permutations
38 FiveMovePermutation =
   ↪ createFiveMovePermutation(AbbreviatedPermutation)
39
40 ## SNIPPET 11 - Omit submoves in permutations
41 FourMovePermutation =
   ↪ createFourMovePermutation(FiveMovePermutation,MergedPermutation)
42
43 ## SNIIPET 12 = Increment count of permutation instances
44
45 count_Abbr_Move =
   ↪ getcounts_Abbr_Move(AbbreviatedPermutation,FiveMovePermutation)
46 count_Abbr = as.data.frame(count_Abbr_Move[1])
47 count_Five = as.data.frame(count_Abbr_Move[2])
48
49 # SNIPPET 13 - Increment linearity count
50 linear_count_mrg = getLinearityCount(MergedPermutation)
51 linear_count_Five = getLinearityCount(FiveMovePermutation)
52 linear_count_Four = getLinearityCount(FourMovePermutation)
53
54
55 # SNIPPET 14 - Increment cyclicity count
56 linear_cycl_mrg = checkCyclicity(linear_count_mrg)
57 linear_cycl_Five = checkCyclicity(linear_count_Five)
58 linear_cycl_Four = checkCyclicity(linear_count_Four)
```

A.17 R script for feature frequency

The following R script harness the functions shown in A.14. The script is created from a series of functions.

```
library(XML)
library(stringr)
library(dplyr)

xml_path = "data/1000/"
xml_files = list.files(paste(getwd(),xml_path, sep='/'),pattern =
 ↪  ".xml")

source("data_funct.R")

##### Feature Frequency #####

#  SNIPPET 1 - Read xml files into dataframe
rawMasterDf = createRawMasterDf(xml_path, xml_files)

#  SNIPPET 2 - Manipulating data (removing unnecessary columns and
 ↪  simplifying terms)
masterDf = createMasterDf(rawMasterDf)

# SNIPPET 5 - feature_submove

# EC Discipline
EC_discipline = masterDf[grep("EC",masterDf$TextFile),]

# WC Discipline
WC_discipline = masterDf[grep("WC",masterDf$TextFile),]

# KDE Discipline
KDE_discipline = masterDf[grep("KDE",masterDf$TextFile),]

# IT Discipline
IT_discipline = masterDf[grep("IT",masterDf$TextFile),]

# IP Discipline
IP_discipline = masterDf[grep("IP",masterDf$TextFile),]

# BOT Discipline
BOT_discipline = masterDf[grep("BOT",masterDf$TextFile),]
```

```
37
38 # IND Discipline
39 IND_discipline = masterDf[grep("IND",masterDf$TextFile),]
40
41 # LING Discipline
42 LING_discipline = masterDf[grep("LING",masterDf$TextFile),]
43
44 # MAT Discipline
45 MAT_discipline = masterDf[grep("MAT",masterDf$TextFile),]
46
47 # MED Discipline
48 MED_discipline = masterDf[grep("MED",masterDf$TextFile),]
49
50 # SNIPPET 6 - : Increment count of feature instances
51 count_EC_discipline = count(EC_discipline,"Features")
52 count_WC_discipline = count(WC_discipline,"Features")
53 count_KDE_discipline = count(KDE_discipline,"Features")
54 count_IT_discipline = count(IT_discipline,"Features")
55 count_IP_discipline = count(IP_discipline,"Features")
56
57 count_BOT_discipline = count(BOT_discipline,"Features")
58 count_IND_discipline = count(IND_discipline,"Features")
59 count_LING_discipline = count(LING_discipline,"Features")
60 count_MAT_discipline = count(MAT_discipline,"Features")
61 count_MED_discipline = count(MED_discipline,"Features")
62
63 count_master = count(masterDf,"Features")
```

A.18 R script for sequence frequency

The following script harnesses the functions shown in A.14. This script is created from a series of functions.

```
library(XML)
library(stringr)
library(dplyr)

xml_path = "data/1000/"
xml_files = list.files(paste(getwd(),xml_path, sep='/'),pattern =
 ↪  ".xml")

source("data_funct.R")

#### Sequence Frequency #####

#  SNIPPET 1 - Read xml files into dataframe
rawMasterDf = createRawMasterDf(xml_path, xml_files)

#  SNIPPET 2 - Manipulating data (removing unnecessary columns and
 ↪  simplifying terms)
masterDf = createMasterDf(rawMasterDf)

# SNIPPET 3 - Create feature only dataframe
feature = masterDf[,-c(2)]

# SNIPPET 4 - Create feature only dataframe
feature_move = createFeatureMove(feature)
feature_submove = createFeatureSubmove(feature)

# Merge feature move and submove dataframes into a single dataframe
MasterPermutation <- cbind(feature_move, Submoves =
 ↪  feature_submove$Submoves)

# SNIPPET 8 - Abbreviate feature permutations
AbbreviatedPermutation = createAbbrPerm(MasterPermutation)

# SNIPPET 9 - Merge sequential identical permutations
MergedPermutation = createMergePerm(AbbreviatedPermutation)

# SNIPPET 10 - Omit submoves in permutations
```

```
35 FiveMovePermutation =
   ↪ createFiveMovePermutation(AbbreviatedPermutation)
36
37 # SNIPPET 11 - Omit submoves in permutations
38 FourMovePermutation =
   ↪ createFourMovePermutation(FiveMovePermutation,MergedPermutation)
39
40 # SNIIPET 12 = Increament count of permutation instances
41
42 count_Abbr_Move =
   ↪ getcounts_Abbr_Move(AbbreviatedPermutation,FiveMovePermutation)
43 count_Abbr = as.data.frame(count_Abbr_Move[1])
44 count_Five = as.data.frame(count_Abbr_Move[2])
45
46 # Tabulate adjacent pairs of moves
47 disciplines <- c("IP", "IT", "KDE", "WC", "EC", "BOT", "IND", "LING",
   ↪ "MAT", "MED")
48 FiveMoveMergedPermutation <- createMergePerm(FiveMovePermutation)
49 MovePermLists <- createMovePermLists(FiveMoveMergedPermutation)
50 DisciplinePairFreqs <- createDisciplinePairFreqsList(MovePermLists,
   ↪ disciplines)
51 DisciplinePairFreqsMat <-
   ↪ createDisciplinePairFreqsMat(DisciplinePairFreqs, disciplines)
52 as.table(DisciplinePairFreqsMat)
53
54 # Create heatmap table for adjacent pairs for each discipline
55 MoveTransMats <- createMoveTransMats(DisciplinePairFreqsMat,
   ↪ disciplines)
56 #
   ↪ https://cran.r-project.org/web/packages/ztable/vignettes/heatmapTable.html
57 library(ztable)
58 library(magrittr)
59 options(ztable.type="html")
60 for (trans_mat_name in names(MoveTransMats)) {
61   trans_mat <- MoveTransMats[[trans_mat_name]]
62   mode(trans_mat) <- "integer"
63   print(makeHeatmap(ztable(trans_mat)), caption=trans_mat_name)
64 }
```

A.19 Standard operating procedure (SOP) for corpus collection

The main priority is accuracy. Good results are only achieved from good data.

Five steps from preparation to double-checking

1. Prepare folder and blank files.
   (a) Create folder for subcorpus, named by two or three letter code, e.g IP
   (b) Create 100 plain text files (txt) within the folder
2. Name each file following the system as exemplified by IP2012dec01_ab.
   (a) The first two or three capital letters indicate the discipline (e.g. IP represents Image Processing).
   (b) The year of the publication is next.
   (c) The month of the issue in lowercase letter follows.
   (d) The two-digit number is the sequential number of the abstract starting from 01. Note the final abstract in each subcorpus is three digits (100).
   (e) The underscore followed by "ab" indicates that the text is an abstract.
3. Go to the website for online journal, e.g. Transactions on Image Processing
   (a) Click on the tab titled “past issues”, then select “2012” and “the first issue”. There are 12 issues and so issue 1 is for January while issue 12 is for December.
   (b) Click on the hyperlink from each title to open the abstract.
   (c) Copy and paste the text of the first abstract into the first txt file and save the file.
   (d) Double check that the first word and the last word of the abstract in the plain text file against the online journal.
   (e) Continue for 99 more abstracts.
4. Double check that each text file contains an abstract and that you have collected all the abstracts. Note: files with size 0kb have no data.

The SOP will alter slightly for journals not housed on IEEE Xplore.

A.20 UAM Corpus Tool Guide

Preparation

1. Install the UAM Corpus Tool
2. Find the location of the rhetorical_moves.xml file and the abstract txt files to be annotated.

Procedure (for the first instance)

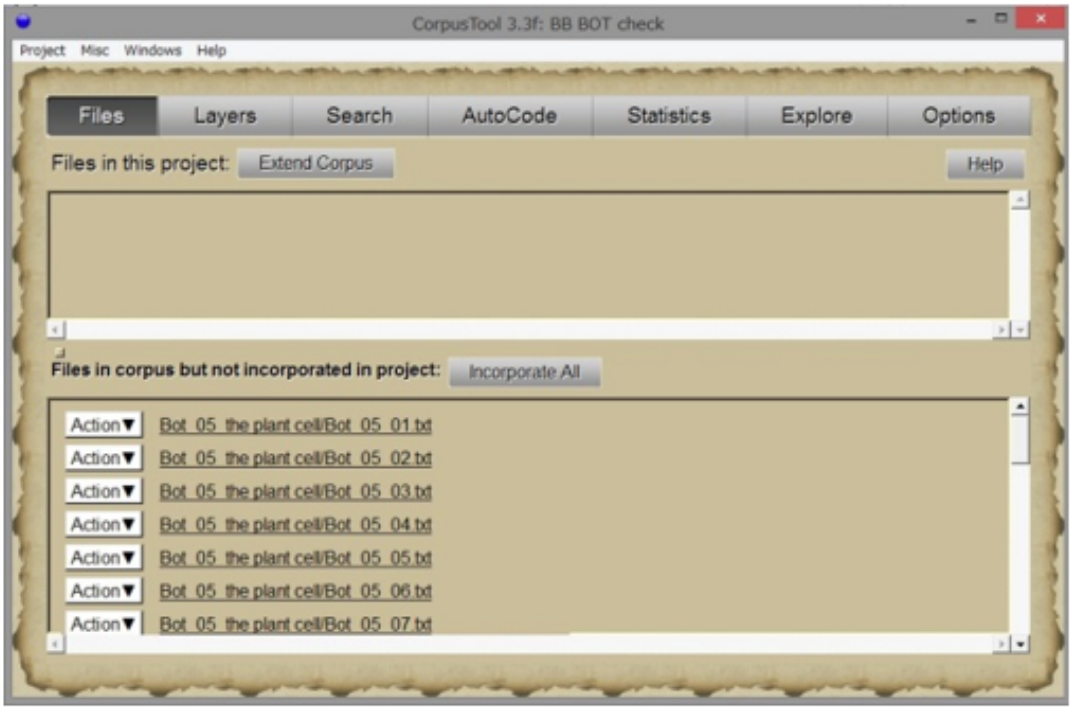
1. Open the UAM Corpus Tool.

Create project

1. Select new project – provide with an appropriate name (e.g. your initials and initials for corpus, BB BOT).

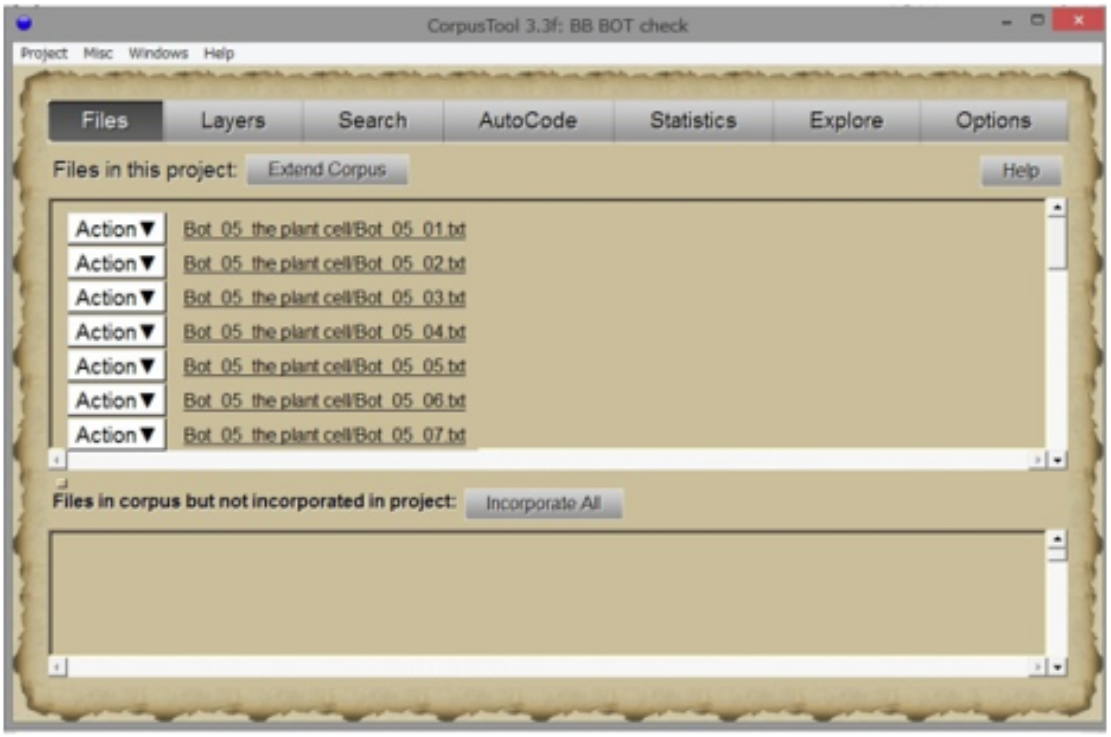
Extend files in corpus

1. Select extend corpus (grey button near top of the screen). A pop-up window will open.
2. Click next.
3. On the screen titled "Corpus Location", click the second radio button "I want to add a folder of text files".
4. Click the white button with three dots [...] and navigate to the folder where the txt files are stored.
5. Click next, and then click finalize.
6. In the bottom section of the window, the files should be displayed



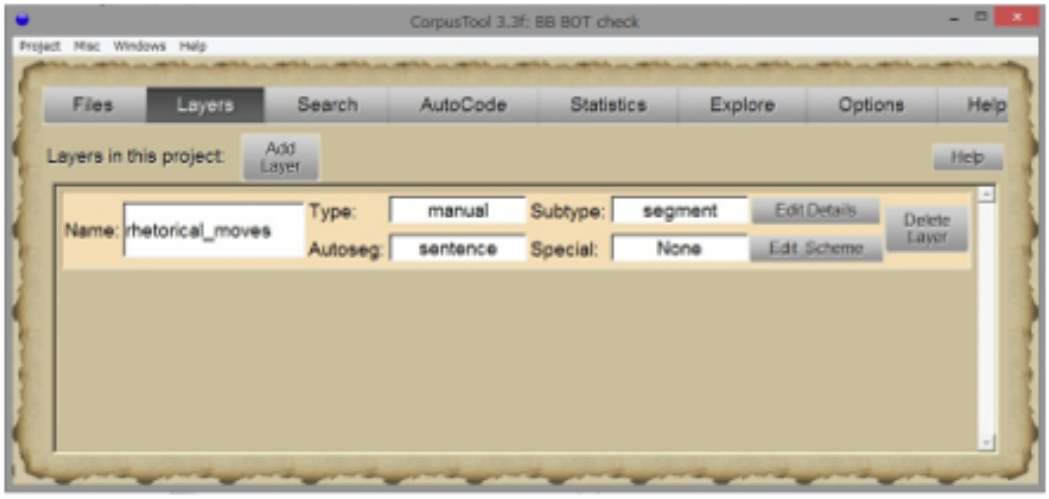
Incorporate files in project

1. Click incorporate all (some windows requiring clicks may open). The files should now be displayed in the top section of the window.

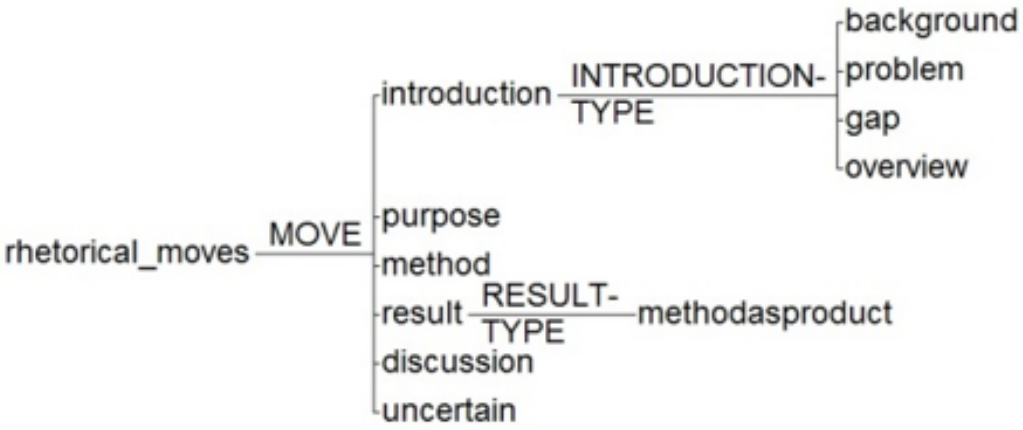


Add scheme

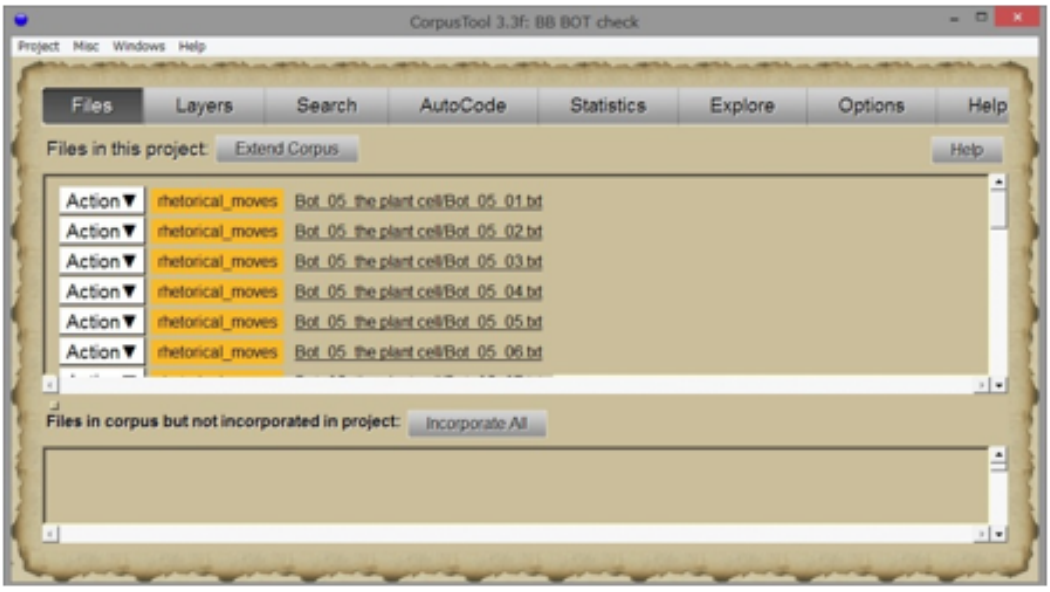
1. Click the Layer button on the top row of grey buttons.
2. Click the Add Layer button.
3. In the pop-up window, click Start.
4. Name the layer "rhetorical_moves"
5. Click Manual annotation.
6. Click Reuse a User-specified Scheme.
7. Click Choose File and navigate to the rhetorical moves.xml file on your computer.
8. Click continue.
9. On the "What kind of segment" screen click Segements within a Document.
10. Click No when asked "Do you need a special layer?".
11. Click Sentences when asked "Should the program automatically segment the text for you?".
12. On the Final check screen, click Create Layer (or cancel or back if you made a mistake).



13. If you click on "Edit scheme" you can view the installed scheme, which should look like this:



14. Click the Files tab and you should see the list of files and a button on the left of each file name with the name of the layer.

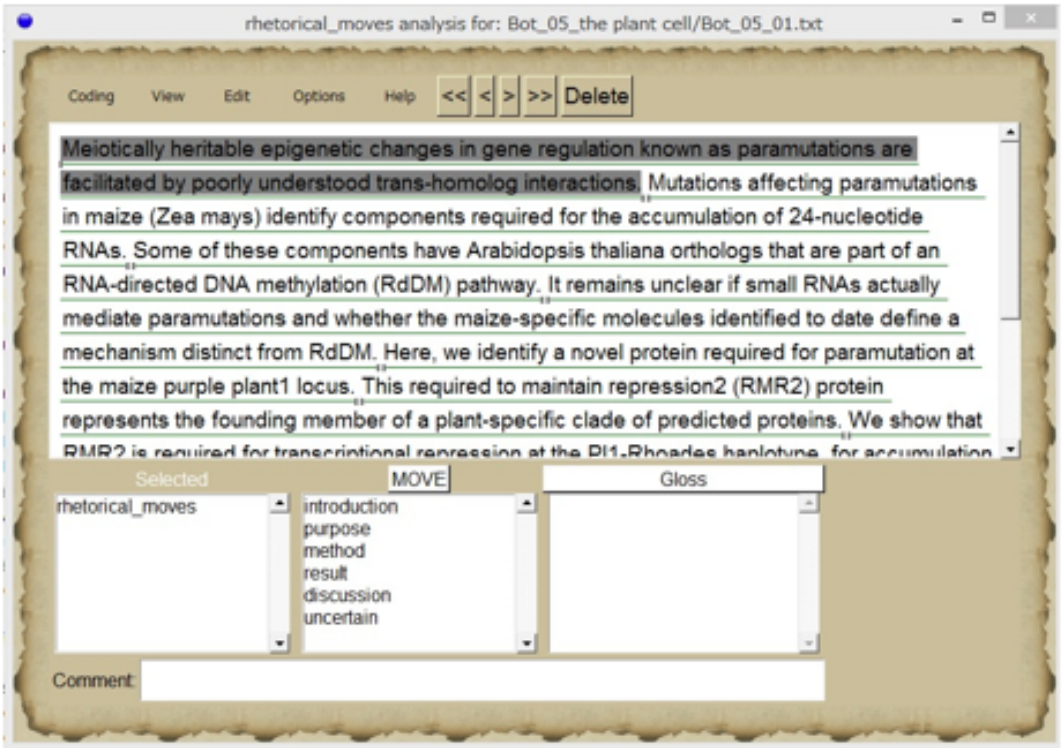


Set auto-advance

1. Click the Options button.
2. On the Coding Options menu, check that Auto-advance on complete is set to True.
3. Click Save Options.

Annotating

1. To annotate click on the Layer button next to the file. (Clicking on the file name will show the text but you cannot annotate it).
2. A pop-up window will appear with the first sentence highlighted.
3. Check all sentences have been identified accurately. If not, you can move the grey blocks at the end of each green line to mark the start and finish points of a sentence. If you need to remove a sentence marker, place the end and start blocks in exactly the same position.
4. When a sentence is highlighted the MOVE choices are displayed at the bottom of the text. Doubleclick on a word to select it. For selected words, doubleclicking will deselect them.



5. Click on the Coding tab in the top left and you can select save. Alternatively, when closing the window, a pop will appear asking whether you want to save. This is a more time-efficient choice.
6. Click on the layer button of the next abstract and continue. Note that the layer button can take three colours that represent annotated, partially-annotated and unannotated.

Bibliography

Abdelmoneim, S. E. (2010). "Plagiarism – What is it? How to avoid it?" In: *14th Alexandia Anaesthesia and Intensive Care Conference*. Alexandria, Egypt. URL: http://www.alexaic.com/alexaicfiles/presentation2010/day3/028001.pdf.

Allwood, J., G.-G. Andersson, L.-G. Andersson, and O. Dahl (1977). *Logic in linguistics*. Cambridge: Cambridge University Press.

Ammon, U. (2011). "Editor's preface". In: *The dominance of English as a language of science: Effects on other languages and language communities*. Ed. by U. Ammon. Berlin, Germany: Walter de Gruyter, p. V. DOI: https://doi.org/10.1515/9783110869484.

Amnuai, W. (2019). "Analyses of rhetorical moves and linguistic realizations in accounting research article abstracts published in international and Thai-based journals". In: *Sage Open* 9.1, pp. 1–9. DOI: https://doi.org/10.1177/2158244018822384.

Anderson, K. and J. Maclean (1997). "A genre analysis study of 80 medical abstracts". In: *Edinburgh working papers in applied linguistics*. Vol. 8. Edinburgh: Edinburgh University Press, pp. 1–23.

Anthony, L. (1998). "Preaching to cannibals: A look at academic writing in the field of engineering". In: *Proceedings of the Japan Conference on English for Specific Purposes*. Ed. by T. Orr, pp. 75–86.

— (1999). "Writing research article introductions in software engineering: How accurate is a standard model?" In: *IEEE Transactions on Professional Communication* 42.1, pp. 38–46. DOI: https://doi.org/10.1109/47.749366.

— (2001). "Characteristic features of research article titles in computer science". In: *IEEE Transactions on Professional Communication* 44.3, pp. 187–194. DOI: https://doi.org/10.1109/47.946464.

— (2013). "A critical look at software tools in corpus linguistics". In: *Linguistic Research* 30.2, pp. 141–161. DOI: https://doi.org/10.17250/khisli.30.2.201308.001.

— (2016). "Looking from the past to the future in ESP through a corpus-based analysis of English for Specific Purposes journal titles". In: *English Teaching & Learning* 40.4. DOI: https://doi.org/10.6330/ETL.2016.40.4.04.

— (2019). *AntConc (Version 3.5.9) [Computer Software]*. Tokyo, Japan. URL: https://www.laurenceanthony.net/software.

Anthony, L. and G. Lashkia (2003). "Mover: A machine learning tool to assist in the reading and writing of technical papers". In: *IEEE: Transactions on Professional*

Communication 46.3, pp. 185–193. DOI: https://doi.org/10.1109/TPC.2003.816789.

Appadurai, A. (1990). "Disjuncture and difference in the global economy". In: *Global culture: Nationalism, globalization and modernity*. Ed. by M. Featherstone. London: Sage, pp. 295–310.

Archer, D.E. (2012). "Corpus annotation: a welcome addition or an interpretation too far?" In: Studies in Variation, Contacts and Change in English eSeries 10. URL: http://www.helsinki.fi/varieng/series/volumes/10/archer/.

Artstein, R. and M. Poesio (2008). "Inter-coder agreement for computational linguistics". In: *Computational Linguistics* 34.4, pp. 555–596. DOI: https://doi.org/10.1162/coli.07-034-R2.

Austin, J. L. (1962). *How to do things with words*. Oxford: Oxford University Press.

Baker, P. (2004). "Querying keywords: Questions of difference, frequency, and sense in keywords analysis". In: *Journal of English Linguistics* 32.4, pp. 346–359. DOI: https://doi.org/10.1177/0075424204269894.

— (2006a). *'The question is, how cruel is it?' Keywords, fox Hunting and the House of Commons*. London. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.180.6013&rep=rep1&type=pdf.

— (2006b). *Using corpora in discourse analysis*. London: Continuum.

Bakhtin, M. M. (1981). "Forms of time and of the chronotope in the novel". In: *The dialogic imagination: Four essays [translated by Caryl Emerson and Michael Holquist]*. Ed. by M. M. Bakhtin. Austin, TX: University of Texas Press, pp. 84–258.

— (1986). "The problem of speech genres and the problem of text in linguistics, philology and the human sciences: An experiment in philosophical analysis". In: *Bakhtin: Speech genres and other late essays*. Ed. by C. Emerson and M Holmquist. TX: Austin: University of Texas Press, pp. 250–317.

Balietti, S., M. Mäs, and D. Helbing (2015). "On disciplinary fragmentation and scientific progress". In: *PloS One* 10.3, e0118747. DOI: https://doi.org/10.1371/journal.pone.0118747.

Barnbrook, G., O. Mason, and R. Krishnamurthy (2013). *Collocation: Applications and implications*. Basingstoke: Palgrave MacMillan.

Baroni, M. and M. Ueyama (2006). "Building general and special purpose corpora by web crawling". In: *Proceedings of the 13th NIJL International Symposium*. URL: http://sslmit.unibo.it/.

Bartholomae, D. (1986). "Inventing the university". In: *Journal of Basic Writing* 5.1, pp. 4–23. URL: https://wac.colostate.edu/jbw/v5n1/bartholomae.pdf.

Barton, D. (2007). *Literacy: An Introduction to the Ecology of Written Language*. 2nd ed. Oxford: Blackwell.

Bawarshi, A. S and M. J. Reiff (2010). *An introduction to history, theory, research, and pedagogy*. Washington, DC: Parlor Press.

Bayerl, P.S. and K.I. Paul (2011). "What determines inter-coder agreement in manual annotations? A meta-analytic investigation". In: *Computational Linguistics* 37.4, pp. 699–725. DOI: http://dx.doi.org/10.1162/COLI_a_00074.

Bazerman, C. (1988). *Shaping Written Knowledge*. Madison: University of Wisconsin Press.

— (1997). "The life of genre, the life in the classroom". In: *Genre and writing: Issues, arguments, alternatives*. Ed. by W. Bishop and H. Ostrom. Portsmouth, NH: Boynton/Cook, pp. 19–26.

Becher, T. and P.R. Trowler (2001). *Academic tribes and territories: intellectual enquiry and the cultures of disciplines*. 2nd ed. Buckingham: Open University Press.

Belcher, D.D. (2006). "English for specific purposes: Teaching to perceived needs and imagined futures in worlds of work, study, and everyday life". In: *TESOL Quarterly* 40.1, pp. 133–156. DOI: https://doi.org/10.2307/40264514.

Belica, C. (1996). "Analysis of temporal changes in corpora". In: *International Journal of Corpus Linguistics* 1.1, pp. 61–73. DOI: https://doi.org/10.1075/ijcl.1.1.05bel.

Benbrahim, M. and K. Ahmad (1995). "Text summarisation: The role of lexical cohesion analysis". In: *The New Review of Document and Text Management* 1, pp. 321–335. URL: https://doc.uments.com/s-text-summarisation-the-role-of-lexical-cohesion-analysis.pdf.

Bender, E. M. and B. Friedman (2018). "Data statements for natural language processing: Toward mitigating system bias and enabling better science". In: *Transactions of the Association for Computational Linguistics* 6, pp. 587–604. URL: https://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00041.

Bereiter, C. and M. Scardamalia (1987). *The Psychology of Written Composition*. New York, NJ: Routledge. DOI: https://doi.org/10.4324/9780203812310.

Berkenkotter, C. and T.N. Huckin (1995). *Genre knowledge in disciplinary communication: Cognition/culture/power*. Hillsdale, NJ: Lawrence Erlbaum.

Berry, E.M. (1981). "The evolution of scientific and medical journals". In: *The New England Journal of Medicine* (305), pp. 400–402. DOI: https://doi.org/10.1056/NEJM198108133050711.

Bhatia, V.K. (1983). "Simplification v. Easification — The case of legal texts 1". In: *Applied linguistics* 4.1, pp. 42–54. DOI: https://doi.org/10.1093/applin/4.1.42.

— (1993). *Analyzing genre: Language use in professional settings*. New York, NY: Longman.

— (1994). "Generic integrity in professional discourse". In: *Text and Talk in Professional Contexts* 6, pp. 61–76.

— (2001). "Analyzing genre: Some conceptual issues". In: *Academic writing in context: Implications and applications (Papers in honour of Tony Dudley-Evans)*. Ed. by M. Hewings. Birmingham, UK: University of Birmingham, pp. 79–92.

— (2004). *Worlds of written discourse: A genre-based view*. London: A&C Black.

Bhatia, V.K. (2013). "The Routledge handbook of discourse analysis". In: ed. by J.P. Gee and M. Handford. London: Routledge, pp. 239–251.

— (2014). *Worlds of written discourse: A genre-based view*. London: Bloomsbury.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511621024.

— (1989). "A typology of English texts". In: *Linguistics* 27.1, pp. 3–44. DOI: https://doi.org/10.1515/ling.1989.27.1.3.

— (1990). "Methodological issues regarding corpus-based analyses of linguistic variation". In: *Literary and Linguistic Computing* 5, pp. 257–69. DOI: https://doi.org/10.1093/llc/5.4.257.

— (1993). "Representativeness in corpus design". In: *Literary and Linguistic Computing* 8.4, pp. 243–257. DOI: https://doi.org/10.1093/llc/8.4.243.

— (1995). *Dimensions of register variation*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511519871.

— (2004). "Historical patterns for the grammatical marking of stance: A cross-register comparison". In: *Journal of Historical Linguistics* 5.1, pp. 107–135. DOI: https://doi.org/10.1075/jhp.5.1.06bib.

Biber, D., U. Connor, and T. A. Upton (2007). *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Amsterdam: John Benjamins. URL: https://journals.openedition.org/asp/925.

Biber, D. and S. Conrad (2009). *Register, genre, and style*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511814358.

Biber, D., S. Conrad, and V. Cortes (2004). "If you look at ...: Lexical bundles in university teaching and textbooks". In: *Applied Linguistics* 25.3, pp. 371–405. DOI: https://doi.org/10.1093/applin/25.3.371.

Biber, D., S. Conrad, and R. Rippen (1998). *Corpus Linguistics*. Cambridge: Foreign Language Teaching and Research Press, Cambridge University Press.

Biber, D. and B. Gray (2010). "Challenging stereotypes about academic writing: Complexity, elaboration, explicitness". In: *Journal of English for Academic Purposes* 9.1, pp. 2–20. DOI: https://doi.org/10.1016/j.jeap.2010.01.001.

— (2013). "Nominalizing the verb phrase in scientific writing". In: *The verb phrase in English: Investigating recent language change with corpora*. Ed. by J. Aarts B.and Close, G. Leech, and S. Wallis. Cambridge: Cambridge University Press, pp. 99–132. DOI: https://doi.org/10.1017/CBO9781139060998.006.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finnegan (1999). *Longman Grammar Spoken and Written English*. Harlow: Longman.

Bird, S. (2006). "NLTK: The natural language toolkit". In: *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, pp. 69–72.

Bird, S., E. Loper, and E. Klein (2009). *Natural language processing with Python*. Sebastopol, CA: O'Reilly Media Inc.

Blake, J. (2014). *Move structure of scientific research abstracts: CARS vs. IMRAD*. Hong Kong Polytechnic University, Hong Kong. URL: https://u-aizu.ac.jp/~jblake/abstracts/CARS_vs_IMRAD_abstract.pdf.

— (2015a). "Incorporating information structure in the EAP curriculum". In: *Conference proceedings of 2nd International Symposium on Innovative Teaching and Research in ESP*. Tokyo: UEC.

— (2015b). "Prescriptive-descriptive disjuncture: Rhetorical organisation of research abstracts in information science". In: *Proceedings of the 8th International Corpus linguistics Conference*. Ed. by F. Formato and A. Hardie. Lancaster University, England, pp. 377–8.

— (2016). "Harnessing keyness: A corpus-based approach to ESP material development". In: *OnCUE* 92, pp. 102–110.

— (2018). "Inter-annotator agreement: By hook or by crook". In: *Proceedings of the Fourth Asia Pacific Corpus Linguistics Conference*. Ed. by Y. Tono and H. Isahara. Takamatsu, Kagawa. Japan, pp. 43–49.

— (2020a). "Automatic identification of tense and grammatical meaning in context". In: *Proceedings of the International Conference on Computers in Education 2020 (Volume II)*. Ed. by H. J. So, M.M. Rodrigo, J. Mason, and A. Mitrovic. Takamatsu, Japan: Asia-Pacific Society for Computers in Education, pp. 739–742. URL: https://apsce.net/upfile/icce2020/ICCE2020-Proceedings-Vol2-FinalUpdated.pdf.

— (2020b). "Development of online tense and aspect identifier for English". In: *CALL for widening participation: Short papers from EuroCALL 2020*. Ed. by K.-M. Frederiksen, S.Larsen, L.Bradley, and S. Thouësny. Research-publishing.net, pp. 1–6. URL: https://doi.org/10.14705/rpnet.2020.48.1161.

— (2020c). "English Verb Analyzer: Identifying tense, voice, aspect, sense and grammatical meaning in context for pedagogic purposes". In: *Extended abstract in 8th Swedish Language Technology Conference 2020 program*. Gothenburg, Sweden: University of Gothenburg. URL: https://gubox.app.box.com/v/SLTC-2020-paper-25.

Blickenstaff, J. and M.J. Moravcsik (1982). "Scientific output in the third world". In: *Scientometrics* 4, pp. 135–169. DOI: https://doi.org/10.1007/BF02018451.

Bochkarev, V. V., E. Y. Lerner, and A.V. Shevlyakova (2014). "Deviations in the Zipf and Heap's laws in natural languages". In: *Journal of Physics. Conference Series* 490.012009, pp. 1–4.

Bojsen-Møller, M., S. Auken, A. J. Devitt, and T. K. Christensen (2020). "Illicit Genres: the case of threatening communications". In: *Sakprosa* 12.1, pp. 1–53. URL: https://journals.uio.no/sakprosa/article/view/7416/7093.

Boleda, G. and S. Evert (2009). *Inter-annotator agreement: Computational lexical semantics (Presentation)*.

Bollen, J., M. A. Rodriquez, and H. Van de Sompel (2006). "Journal status". In: *Scientometrics* 69.3, pp. 669–687. URL: https://link.springer.com/content/pdf/10.1007/s11192-006-0176-z.pdf.

Bourdieu, P., J.-C. Passeron, and M. de Saint Martin (1996). *Academic discourse: Linguistic misunderstanding and professorial power*. Stanford, CA: Stanford University Press.

Bowker, L. and J. Pearson (2002). *Working with specialized language: A practical guide to using corpora*. London: Routledge. DOI: https://doi.org/10.4324/9780203469255.

Bredbenner, K. and S.M. Simon (2019). "Video abstracts and plain language summaries are more effective than graphical abstracts and published abstracts". In: *PloS one* 14.11, e0224697. DOI: https://doi.org/10.1371/journal.pone.0224697.

Brembs, B. (2019). "Reliable novelty: New should not trump true". In: *PLoS Biology* 17.2, e3000117.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781316410899.

Britannica (2013). *Encyclopedia Britannica [Ultimate edition]*. Rugeley: Focus Multimedia Ltd.

Brysbaert, M., P. Mandera, and E. Keuleers (2017). "Corpus Linguistics". In: *Research methods in psycholinguistics and the neurobiology of language: A practical guide*. Ed. by A. M. B. de Groot and P. Hagoort. Hoboken, NJ: Wiley-Blackwell, pp. 230–246.

Buchstaller, I. and G. Khattab (2013). "Population samples". In: *Research methods in linguistics*. Ed. by R. J. Podesva and D. Sharma. Cambridge: Cambridge University Press, pp. 74–95.

Bunton, D. (2014). "Generic moves in Ph.D. thesis introductions". In: *Academic discourse*. Ed. by J. Flowerdew. London: Routledge, pp. 67–85. DOI: https://doi.org/10.4324/9781315838069.

Burton, L. L. (2004). *Mathematicians as enquirers: Learning about learning mathematics*. Vol. 34. Dordrecht, Netherlands: Springer Science & Business Media. DOI: https://doi.org/10.1007/978-1-4020-7908-5_8.

Can, S., E.Karabacak, and J. Qin (2016). "Structure of moves in research article abstracts in applied linguistics". In: *Publications* 4.3, p. 23. DOI: https://doi.org/10.3390/publications4030023.

Canagarajah, A. S. (1996). "'Nondiscursive' requirements in academic publishing, material resources of periphery scholars, and the politics of knowledge production". In: *Written Communication* 13.4, pp. 435–472. DOI: https://doi.org/10.1177/0741088396013004001.

Cao, Y. and R. Xiao (2013). "A multi-dimensional contrastive study of English abstracts by native and non-native writers". In: *Corpora* 8.2, pp. 209–234. DOI: https://doi.org/10.3366/cor.2013.0041.

Carter, R. and M. McCarthy (2001). "Size isn't everything: Spoken English, corpus and the classroom". In: *TESOL Quarterly* 35.2, pp. 337–340. DOI: https://doi.org/10.2307/3587654.

— (2006). *Cambridge grammar of English: A comprehensive guide; spoken and written English grammar and usage*. Cambridge: Cambridge University Press.

Casanave, C. P. (1998). "Transitions: The balancing act of bilingual academics". In: *Journal of Second Language Writing* 7.2, pp. 175–203. DOI: https://doi.org/10.1016/S1060-3743(98)90012-1.

Chapman, C. A., J. C.r Bicca-Marques, S. Calvignac-Spencer, P. Fan, P. J. Fashing, J. Gogarten, S. Guo, C. A. Hemingway, F. Leendertz, B. Li, et al. (2019). "Games academics play and their consequences: How authorship, h-index and journal impact factors are shaping the future of academia". In: *Proceedings of the Royal Society B* 286.1916. DOI: https://doi.org/10.1098/rspb.2019.2047.

Charniak, E. (1997). "Statistical techniques for natural language parsing". In: *AI Magazine* 18.4, p. 33. DOI: https://doi.org/10.1609/aimag.v18i4.1320.

Cheng, W. (2012). *Exploring corpus linguistics: Language in action*. Abingdon, Oxon: Routledge.

Chiswick, B. R. and P. W. Miller (2005). "Linguistic distance: A quantitative measure of the distance between English and other languages". In: *Journal of Multilingual and Multicultural Development* 26.1, pp. 1–11. DOI: https://doi.org/10.1080/14790710508668395.

Chomsky, N. (2002). *Syntactic structures*. The Hague, Netherlands: Walter de Gruyter.

Christenhusz, M.J.M. and J. W. Byng (2016). "The number of known plants species in the world and its annual increase". In: *Phytotaxa* 261.3, pp. 201–217. DOI: https://doi.org/10.11646/phytotaxa.261.3.1.

Church, K.W. and W.A. Gale (1994). "What's wrong with adding one". In: *Corpus linguistics: Readings in a widening discipline*. Ed. by G. Sampson and D. McCarthy. London, pp. 95–102.

— (1995). "Poisson mixtures". In: *Journal of Natural Language Engineering* 1.2, pp. 163–190. DOI: https://doi.org/10.1017/S1351324900000139.

Clancy, B. (2012). "Building for a variety of a language". In: *The Routledge handbook of corpus linguistics*. Ed. by Anne O'Keeffe and Michael McCarthy. Abingdon, Oxon: Routledge, pp. 80–92.

Clear, J. (1992). "Corpus sampling". In: *New directions in English language corpora*. Ed. by G. Leitner. 21st ed. New York, NY: Mouton de Gruyter.

Clyne, M. (1987). "Cultural differences in the organization of academic texts: English and German". In: *Journal of Pragmatics* 11.2, pp. 211–241. DOI: https://doi.org/10.1016/0378-2166(87)90196-2.

— (1991). "The sociocultural dimension: the dilemma of the German-speaking scholar". In: *Subject-oriented texts. Languages for special purposes and text theory*. Ed. by Hartmut Schröder. Berlin: de Gruyter, pp. 49–68.

Coe, R. M. (2002). "The new rhetoric of genre: Writing political briefs". In: *Genre in the classroom: Multiple perspectives*. Ed. by A.M. Johns. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 197–207.

Cohen, J. (1960). "A coefficient of agreement for nominal scales". In: *Education and psychological measurement*. Vol. 20, pp. 37–46.

Connor, U. and A. Mauranen (1999). "Linguistic analysis of grant proposals: European Union research grants". In: *English for Specific Purposes* 18.1, pp. 47–62. DOI: https://doi.org/10.1016/S0889-4906(97)00026-4.

Cotos, E. and N. Pendar (2016). "Discourse classification into rhetorical functions for AWE feedback". In: *CALICO Journal* 33.1, pp. 92–116. DOI: https://doi.org/10.1558/cj.v33i1.27047.

Coulthard, M. and A. Johnson (2007). *An introduction to forensic linguistics*. London/New York: Routledge.

Coxhead, A. (2000). "A new academic word list". In: *TESOL Quarterly* 34.2, pp. 213–238. DOI: https://doi.org/10.2307/3587951.

Crawley, M.J. (2007). *The R book*. Chichester: John Wiley & Sons, Ltd.

Crookes, G. (1984). "Towards a validated analysis of scientific text structure". In: *Applied Linguistics* 7.1, pp. 57–70. DOI: https://doi.org/10.1093/applin/7.1.57.

Cross, C. and C. Oppenheim (2006). "A genre analysis of scientific abstracts". In: *Journal of Documentation* 62.4, pp. 428–446. DOI: https://doi.org/10.1108/00220410610700953.

Cunningham, H., D. Maynard, and K. Bontcheva (2011). *Text processing with GATE*. Murphys, CA: Gateway Press.

Dalpanagioti, T. (2018). "A frame-semantic approach to co-occurrence patterns: A lexicographic study of English and Greek motion verbs". In: *International Journal of Lexicography* 31.4, pp. 420–451.

Davies, M. (2007). *TIME Magazine Corpus: 100 million words, 1920s-2000s*. URL: https://www.english-corpora.org/time/.

Day, R.A. and B. Gastel (2006). *How to write and publish scientific papers*. 6th ed. Cambridge: Cambridge University.

Dayrell, C., A. Candido Jr., G. Lima, D. Machado Jr., A. Copestake, V. D. Feltrim, S. Tagnin, and S. Aluisio (2012). "Rhetorical move detection in English abstracts: multi-label sentence classifiers and their annotated corpora". In: *Proceedings of 8th Language Resources Evaluation Conference*. Istanbul, Turkey. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/734_Paper.pdf.

De Jonge, E. and M. van der Loo (2013). *An introduction to data cleaning with R*. The Hague, Netherlands: Statistics Netherlands.

Derewianka, B. (2003). ""Technicality," from Science and Language Links: Classroom Implications". In: *Building academic literacy: An anthology for reading apprenticeship*. Ed. by A. Fielding and R. Schoenbach. San Franscisco, CA: Jossey-Bass, pp. 253–258.

Devitt, A.J. (2004). *Writing genres*. Carbondale, IL: Southern Illinois University.

Di Bitetti, M.S. and J. A. Ferreras (2017). "Publish (in English) or perish: The effect on citation rate of using languages other than English in scientific publications". In: *Ambio* 46.1, pp. 121–127. DOI: https://doi.org/10.1007/s13280-016-0820-7.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative and mixed methodologies*. Oxford: Oxford University Press.

Doró, K. (2013). "The rhetoric structure of research article abstracts in English studies journals". In: *Prague Journal of English Studies* 2.1, pp. 119–139. DOI: https://doi.org/10.2478/pjes-2014-0013.

dos Santos, M.B. (1996). "The textual organisation of research paper abstracts in applied linguistics". In: *Text* 16, pp. 481–499. DOI: https://doi.org/10.1515/text.1.1996.16.4.481.

Drew, P. (2004). "Integrating qualitative analysis of evaluative discourse with the quantitative approach of corpus linguistics". In: *Strategies in academic discourse*. Ed. by E. Tognini-Bonelli and G. Del Lungo Camiciotti. Amsterdam: John Benjamins Publishing, pp. 217–229.

Dudley-Evans, T. (2000). "Genre analysis: A key to a theory of ESP?" In: *Ibérica* 2, pp. 3–11. URL: http://www.aelfe.org/documents/text2-Dudley.pdf.

— (2002). "Genre analysis: An approach to text analysis for ESP". In: *Advances in written text analysis*. Ed. by M. Coulthard. London: Routledge, pp. 233–242.

Dunn, J. (2019). "Global syntactic variation in seven languages: Toward a computational dialectology". In: *Frontiers in Artificial Intelligence* 2, p. 15. DOI: https://doi.org/10.3389/frai.2019.00015.

Eberhard, D. M., G. F. Simons, and C.D. Fennig, eds. (2019). 22nd ed. Dallas, TX: SIL International. URL: https://www.ethnologue.com/.

Enard, W. and P. Svante (2009). *Probing the evolution of human language in a model organism*. URL: https://www.youtube.com/watch?v=k27DfgKGVp8.

Englander, K. (2013). *Writing and publishing science research papers in English: A global perspective*. New York: Springer Science & Business Media.

Enk, A. van and K. Power (2017). "What is a research article?: Genre variability and data selection in genre research". In: *Journal of English for Academic Purposes* 29, pp. 1–11. DOI: https://doi.org/10.1016/j.jeap.2017.07.002.

Esfandiari, R. (2014). "Realization of rhetorical moves and verb tense variation in two subdisciplines of computer sciences: Artificial intelligence and architecture". In: *International Journal of Language Learning and Applied Linguistics World* 5 (2), pp. 564–573.

Evert, S. (2009). "Corpora and collocations". In: *Corpus linguistics. An international handbook*. Ed. by Anke Lüdeling and Merja Kytö. Vol. 2. Berlin, Germany: De Gruyter Mouton, pp. 1212–1248.

Faigley, L. (1986). "Competing theories of process: A critique and a proposal". In: *College Composition and Communication* 48 (6), pp. 527–542. DOI: https://doi.org/10.2307/376707.

Fairclough, N. (1992). *Discourse and social change*. Cambridge: Polity Press.

— (2003). *Analysing discourse: Textual analysis for social research*. London: Routledge.

Farley, P.C. (2017). "Genre analysis of decision letters from editors of scientific journals: Building on Flowerdew and Dudley-Evans (2002)". In: *Applied Linguistics* 38.6, pp. 896–905. DOI: https://doi.org/10.1093/applin/amw043.

Feak, C.B., S.M. Reinhart, and A. Sinsheimer (2000). "A preliminary analysis of law review notes". In: *English for Specific Purposes* 19.3, pp. 197–220. DOI: https://doi.org/10.1016/S0889-4906(99)00007-1.

Ferguson, G. (2007). "The global spread of English, scientific communication and ESP: Questions of equity, access and domain loss". In: *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE)* 13, pp. 7–38.

Fillmore, C. J. (1992). "Corpus Linguistics' or 'Computer-aided armchair linguistics'". In: *Directions in corpus linguistics. Proceedings of Nobel Symposium*. Vol. 82, pp. 35–60.

Firth, J.R. (1957). "Modes of meaning". In: *Papers in linguistics, 1934-1951*. Ed. by J.R. Firth. London: Oxford University.

Flower, L. (1979). "Writer-based prose: A cognitive basis for problems in writing". In: *College English* 41 (1), pp. 19–37. DOI: https://doi.org/10.2307/376357.

Flowerdew, J. (1996). "Concordancing in language learning". In: *The power of CALL*. Ed. by Martha Carswell Pennington. Houston, TX: Athelstan, pp. 97–113.

— (2000). "Discourse community, legitimate peripheral participation, and the nonnative-English speaking scholar". In: *TESOL Quarterly* 34, pp. 1–127. DOI: https://doi.org/10.2307/3588099.

— (2002). "Genre in the classroom: A linguistic approach". In: *Genre in the classroom: Multiple perspectives*. Ed. by A. M. Johns. Mahwah, NJ: Lawrence Erlbaum, pp. 91–104.

— (2008). "Scholarly writers who use English as an additional language: What can Goffman's "Stigma" tell us?" In: *Journal of English for Academic Purposes* 7.2, pp. 77–86. DOI: https://doi.org/10.1016/j.jeap.2008.03.002.

Flowerdew, L. (1998). "Corpus linguistic techniques applied to text linguistics". In: *System* 26.4, pp. 541–542. DOI: https://doi.org/10.1016/S0346-251X(98)00039-6.

— (2004). "The argument for using English specialized corpora to understand academic and professional settings". In: *Discourse in the professions: Perspectives from corpus linguistics*. Ed. by T.A. Upton and U. Connor. Amsterdam: John Benjamins, pp. 11–33.

— (2009). "Applying corpus linguistics to pedagogy: A critical evaluation". In: *International Journal of Corpus Linguistics* 14.3, pp. 393–417. DOI: https://doi.org/10.1075/ijcl.14.3.05flo.

— (2016). "A genre-inspired and lexico-grammatical approach for helping postgraduate students craft research grant proposals". In: *English for Specific Purposes* 42, pp. 1–12. DOI: https://doi.org/10.1016/j.esp.2015.10.001.

Fort, K., A. Nazarenko, and S. Rosset (2012). "Modeling the complexity of manual annotation tasks: A grid of analysis". In: *Proceedings of the International Conference on Computational Linguistics (COLING 2012)*, pp. 895–910.

Francis, W. N. (1965). "A standard corpus of edited present-day American English". In: *College English* 26.4, pp. 267–273.

Frow, J. (2014). *Genre*. 2nd. Abingdon, Oxon: Routledge.

Fuller, G. (2005). "Reading science: Critical and functional perspectives on discourses of science". In: *Cultivating science*. Ed. by J.R. Martin and R. Veel. Abingdon, Oxon: Routledge, pp. 35–62.

Galasinski, D. (2019). *Discourses of men's suicide notes: A qualitative analysis*. London: Bloomsbury Publishing.

Galtung, J. (1979). "Deductive thinking and political practice. An essay of the teutonic intellectual style". In: *Papers on methodology, Essays on methodology*. Ed. by J. Galtung. Vol. II. Copenhagen.

Garside, R., G. Leech, and T. McEnery (1997). *Corpus annotation: Linguistic information from computer text corpora*. Abingdon, Oxon: Taylor and Francis.

Garvey, W.D. (1979). *Communication: The Essence of Science*. Oxford: Pergamon.

Geiger, R.S., K. Yu, Y. Yang, M. Dai, J. Qiu, R. Tang, and J. Huang (2020). "Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?" In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 325–336. URL: https://dl.acm.org/doi/pdf/10.1145/3351095.3372862.

Gelbukh, A. (2011). "Part-of-speech tagging from 97% to 100%: Is it time for some linguistics?" In: *Computational Linguistics and Intelligent Text Processing, 12th International Conference, CICLing 2011, Proceedings, Part I. Lecture Notes in Computer Science 6608*. Ed. by C.D. Manning. Springer, pp. 171–189.

Georgiev, H. (2006). *English algorithmic grammar*. Continuum.

Gledhill, C. (2009). "Colligation and the cohesive function of present and past tense in the scientific research article". In: *Les temps et les textes de spécialité*. Ed. by D. Banks. Paris: l'Harmattan, pp. 65–84. URL: https://silo.tips/download/christopher-gledhill-universite-marc-bloch-strasbourg.

Goh, G.-Y. (2011). "Choosing a reference corpus for keyword calculation". In: *Linguistic Research* 28.1, pp. 239–256. URL: https://pdfs.semanticscholar.org/8900/5d84dcffd802182c69eefd84edaa00633ba7.pdf.

Gopen, G.D. and J.A. Swan (1990). *The science of scientific writing*. Vol. 78. 6, pp. 550–558. URL: https://www.americanscientist.org/blog/the-long-view/the-science-of-scientific-writing.

Graddol, D. (1997). *The Future of English? British Council*. URL: http://www.britishcouncil.org/learning-elt-future.pdf.

Graetz, N. (1982). "Teaching EFL students to extract structural information from abstracts". In: *Proceedings of International Symposium on Languages for Specific Purposes*. Eindhoven, The Netherlands: ERIC, pp. 1–22. URL: https://files.eric.ed.gov/fulltext/ED224327.pdf.

Granger, S. (2009). "Commentary on part I: Learner corpora: A window onto the L2 phrasicon". In: *Researching collocations in another language*. Springer, pp. 60–65.

Gribbin, J. (2011). *In search of Schrödinger's cat: Quantum physics and reality*. London: Bantam.

Gries, S.Th. (2006). "Some proposals towards more rigorous corpus linguistics". In: *Zeitschrift für Anglistik und Amerikanistik* 54.2, pp. 191–202.

— (2008). "Dispersions and adjusted frequencies in corpora". In: *International Journal of Corpus Linguistics* 13.4, pp. 403–437.

— (2011). "Commentary". In: *Current methods in historical semantics*. Ed. by K. Allan & J.Robinson. Berlin & New York: Mouton de Gruyter, pp. 184–195.

— (2013). *Statistics for linguistics with R. A practical introduction (2nd Ed.)* Boston: De Gruyter Mouton.

— (2015). "Some current quantitative problems in corpus linguistics and a sketch of some solutions". In: *Language and Linguistics* 16.1, pp. 93–117. DOI: https://doi.org/10.1177/1606822X14556606.

— (2020). "Analyzing dispersion". In: ed. by M. Paquot and S.Th. Gries. Dordrecht, The Netherlands: Springer, pp. 99–118.

Gries, S.Th. and A.L. Berez (2017). "Linguistic annotation in/for corpus linguistics". In: *Handbook of linguistic annotation*. Ed. by N. Ide and J. Pustejovsky. Dordrecht, The Netherlands: Springer, pp. 379–409.

Gries, S.Th. and J. Newman (2013). "Creating and using corpora". In: *Research methods in linguistics*. Ed. by R.J. Podesva and D. Sharma. Cambridge: Cambridge University Press, pp. 257–287.

Gries, S.Th. and A. Stefanowitsch (2004). "Extending collostructional analysis: A corpus-based perspective on alternations". In: *International Journal of Corpus Linguistics* 9.1, pp. 97–129. DOI: https://doi.org/10.1075/ijcl.9.1.06gri.

Grieve, J., D. Biber, E. Friginal, and T. Nekrasova (2010). "Variation among blog text types: A multi-dimensional analysis". In: *Genres on the web: Corpus studies and computational models*. Ed. by A. Mehler, S. Sharoff, and M. Santini. New York: Springer-Verlag.

Gross, A. G. (1996). *The rhetoric of science*. Cambridge, MA: Harvard University Press.

Groves, T. and K. Abbasi (2004). "Screening research papers by reading abstracts". In: *British Medical Journal* 329.7464, pp. 470–471. DOI: https://dx.doi.org/10.1136/bmj.329.7464.470.

Gu, X. and K. Blackmore (2017). "Characterisation of academic journals in the digital age". In: *Scientometrics* 110.3, pp. 1333–1350. DOI: https://dx.doi.org/10.1007/s11192-016-2219-4.

Gunning, R. (1969). "The fog index after twenty years". In: *Journal of Business Communication* 6.2, pp. 3–13. DOI: https://doi.org/10.1177/002194366900600202.

Gurevych, I. and Y. Miyao, eds. (2018). *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics. URL: http://aclweb.org/anthology/.

Haan, P. de and R. van Hout (1986). "Statistics and corpus analysis". In: *Corpus Linguistics II*. Ed. by J. Aarts and W. Meijs. Amsterdam: Rodopi, pp. 79–97.

Haggan, M. (2004). "Research paper titles in literature, linguistics and science: Dimensions of attraction". In: *Journal of Pragmatics* 36.2, pp. 293–317. DOI: https://doi.org/10.1016/S0378-2166(03)00090-0.

Halliday, M.A.K. (1985). *An introduction to functional grammar*. 2nd. London: Edward Arnold.

— (1991). "Corpus studies and probabilistic grammar". In: *English corpus linguistics: Studies in Honour of Jansvartvik*. Ed. by K. Aijmer and B. Altenberg. London: Longman, pp. 30–40. DOI: https://doi.org/10.4324/9781315845890.

— (1992). "Language as system and language as instance: The corpus as a theoretical construct". In: *Directions in corpus linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4—8 August 1991*, pp. 61–77.

— (1994). *An introduction to functional grammar*. 2nd. London: Edward Arnold.

Halliday, M.A.K. and J.R. Martin (1993). *Writing science: Literacy and discursive power*. London: Falmer Press.

Halliday, M.A.K. and C.M.I.M. Matthiessen (2004). *An introduction to functional grammar*. 3rd. London: Routledge.

Hanauer, D.I. and K. Englander (2013). *Scientific writing in a second language*. Anderson, SC: Parlor Press.

Hanauer, D.I., C.L. Sheridan, and K. Englander (2019). "Linguistic injustice in the writing of research articles in English as a second language: Data from Taiwanese and Mexican researchers". In: *Written Communication* 36.1, pp. 136–154. DOI: https://doi.org/10.1177/0741088318804821.

Hancioğlu, N., S. Neufeld, and J. Eldridge (2008). "Through the looking glass and into the land of lexico-grammar". In: *English for Specific Purposes* 27.4, pp. 459–479. DOI: https://doi.org/10.1016/j.esp.2008.08.001.

Handford, M.J.A. (2007). "The genre of the business meeting: A corpus-based study". PhD thesis. University of Nottingham.

— (2012a). "Professional communication and corpus linguistics". In: *Corpus applications in applied linguistics*. Ed. by K. Hyland, M. H. Chau, and M.J.A Handford. London: Bloomsbury, pp. 13–29.

— (2012b). "What can a corpus tell us about specialist genres". In: *The Routledge handbook of corpus linguistics*. Ed. by A. O'Keeffe and M. McCarthy. London: Routledge, pp. 255–269.

Hardie, A. (2007). "From legacy encodings to Unicode: the graphical and logical principles in the scripts of South Asia". In: *Language Resources and Evaluation* 41.1, pp. 1–25. URL: https://www.jstor.org/stable/30200570.

Hartley, J. (1994). "Three ways to improve the clarity of journal abstracts". In: *British Journal of Educational Psychology* 64.2, pp. 331–343. DOI: https://doi.org/10.1111/j.2044-8279.1994.tb01106.x.

Hartley, J. (2003). "Improving the clarity of journal abstracts in psychology: The case for structure". In: *Science Communication* 24.3, pp. 366–379. DOI: https://doi.org/10.1177/1075547002250301.

— (2007). "There's more to the title than meets the eye: Exploring the possibilities". In: *Journal of Technical Writing and Communication* 37.1, pp. 95–101. DOI: https://doi.org/10.2190/BJ16-8385-7Q73-1162.

— (2014). "Current findings from research on structured abstracts: an update". In: *Journal of the Medical Library Association* 102.3, p. 146. DOI: https://dx.doi.org/10.3163\%2F1536-5050.102.3.002.

Hartley, J. and M. Benjamin (1998). "An evaluation of structured abstracts in journals published by the British psychological society". In: *British Journal of Educational Psychology* 68, pp. 443–456. DOI: http://dx.doi.org/10.1111/j.2044-8279.1998.tb01303.x.

Harwood, N. (2005). "'Nowhere has anyone attempted... In this article I aim to do just that': A corpus-based study of self-promotional I and we in academic writing across four disciplines". In: *Journal of Pragmatics* 37.8, pp. 1207–1231. DOI: https://doi.org/10.1016/j.pragma.2005.01.012.

Hasan, R. (1996). "What's going on: A dynamic view of context in language". In: *Ways of saying: Ways of meaning: Selected papers of Ruqaiya Hasan*. Ed. by C. Cloran, D. Butt, and G. Williams. London: Cassell, pp. 37–50.

Haswell, R.H. (1991). *Gaining ground in college writing: Tales of development and interpretation*. Dallas, TX: Southern Methodist University Press.

Hatim, B. and I. Mason (1997). *The translator as communicator*. London: Routledge.

Hayashi, T., K. Tomioka, and O. Yonemitsu (1998). *Asymmetric synthesis: Graphical abstracts and experimental methods*. Tokyo, Japan: Kodansha.

Hayer, C.-A., M. Kaemingk, J.J. Breeggemann, D. Dembkowski, D. Deslauriers, and T. Rapp (2013). "Pressures to publish: Catalysts for the loss of scientific writing integrity?" In: *Fisheries* 38 (8), pp. 348–351. DOI: https://doi.org/10.1080/03632415.2013.813845.

Haynes, R.B., C.D. Mulrow, E.J. Huth, D.G. Altman, and M.J. Gardner (1990). "More informative abstracts revisited". In: *Annals of Internal Medicine* 113.1, pp. 69–76.

Hayward, R.S.A., M.C. Wilson, S.R. Tunis, E.B. Bass, H.R. Rubin, and R.B. Haynes (1993). "More informative abstracts of articles describing clinical practice guidelines". In: *Annals of Internal Medicine* 118.9, pp. 731–737. DOI: https://doi.org/10.7326/0003-4819-118-9-199305010-00012.

Heap, B.R. (1963). "Permutations by interchanges". In: *The Computer Journal* 6.3, pp. 293–298. DOI: https://doi.org/10.1093/comjnl/6.3.293.

Henry, A. and R.L. Roseberry (2001). "A narrow-angled corpus analysis of moves and strategies of the genre: 'Letter of Application'". In: *English for Specific Purposes* 20.2, pp. 153–167. DOI: https://doi.org/10.1016/S0889-4906(99)00037-X.

Hinds, J. (1987). "Reader versus writer responsibility: A new typology". In: *Writing across languages: Analysis of L2 texts*. Ed. by U.Connor and R.B. Kaplan. Reading, MA: Addison-Wesley, pp. 141–152.

Hirsch, J.E. (2010). "An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship". In: *Scientometrics* 85.3, pp. 741–754.

Hoey, M. (1983). *On the surface of discourse*. London: Allen & Unwin.

— (2001). *Textual interaction: An introduction to written text analysis*. London: Routledge.

— (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

Hoffmann, A.H. (2010). *Scientific writing and communication: Papers, proposals and communication*. Oxford: Oxford University Press.

Holes, C. (1995). *Modern Arabic: Structures, functions and varieties*. London: Longman.

Holmes, R. (1997). "Genre analysis, and the social sciences: An investigation of the structure of research article discussion sections in three disciplines". In: *English for Specific Purposes* 16.4, pp. 321–337. DOI: https://doi.org/10.1016/S0889-4906(96)00038-5.

Holtz, M. (2009). "Nominalization in scientific discourse: A corpus-based study of abstracts and research articles". In: *Proceedings of the 5th Corpus Linguistics Conference*. Ed. by M. Mahlberg, V. González-Díaz, and C. Smith. Liverpool, UK.

— (2011). "Lexico-grammatical properties of abstracts and research articles. A corpus-based study of scientific discourse from multiple disciplines". PhD thesis. Technische Universität. URL: http://tuprints.ulb.tu-darmstadt.de/2638/.

Hovy, E. and J. Lavid (2010). "Towards a 'science'of corpus annotation: A new methodological challenge for corpus linguistics". In: *International Journal of Translation* 22.1, pp. 13–36. URL: https://www.cs.cmu.edu/~hovy/papers/10KNS-annotation-Hovy-Lavid.pdf.

Huckin, T.N. (2001). "Abstracting from abstracts". In: *Academic writing in context: Implications and applications*. Ed. by M. Hewings. Birmingham: University of Birmingham, pp. 93–103.

Huckin, T.N. and L.A. Olsen (1984). "The need for professionally oriented ESL instruction in the United States". In: *TESOL Quarterly* 18.2, pp. 273–294. DOI: https://doi.org/10.2307/3586694.

Hunston, S. (1994). "Evaluation and organization in a sample of written academic discourse". In: *Advances in written text analysis*. Ed. by M. Coulthard. London: Routledge, pp. 191–218.

— (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

— (2013). *Flavours of corpus linguistics*. Birmingham: University of Birmingham.

Hunston, S. and G. Francis (1999). *Pattern grammar: A corpus driven approach to the lexical grammar of English*. Amsterdam, Netherlands: John Benjamins.

Hyland, K. (2002). *Teaching and researching writing*. Harlow: Pearson Education.

Hyland, K. (2003). "Genre-based pedagogies: A social response to process". In: *Journal of Second Language Writing* 12 (1), pp. 17–29. DOI: https://doi.org/10.1016/S1060-3743(02)00124-8.

— (2004). *Disciplinary discourses: Social interactions in academic writing*. Ann Arbor, MI: University of Michigan.

— (2005a). "Digging up texts and transcripts: Confessions of a discourse analyst". In: *Second language writing research: Perspectives on the process of knowledge construction*. Ed. by P. K. Matsuda and T. Silva. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 177–189. DOI: https://doi.org/10.4324/9781410612755.

— (2005b). *Metadiscourse: Exploring interaction in writing metadiscourse*. London: Continuum.

— (2006). *English for Academic Purposes: An advanced resource book*. New York, NY: Routledge.

— (2007). "Genre pedagogy: Language, literacy and L2 writing instruction". In: *Journal of Second Language Writing* 16 (3), pp. 148–164. DOI: https://doi.org/10.1016/j.jslw.2007.07.005.

— (2008). "As can be seen: Lexical bundles and disciplinary variation". In: *English for Specific Purposes* 27.1, pp. 4–21. DOI: https://doi.org/10.1016/j.esp.2007.06.001.

— (2009). *Academic discourse: English in a global context*. London: Continuum.

— (2010). "English for professional academic purposes: Writing for scholarly publication". In: *English for specific purposes in theory and practice*. Ed. by D. Belcher. Ann Arbor, MI: University of Michigan, pp. 83–105.

— (2012a). "Corpora and academic discourse". In: *Corpus applications in applied linguistics*. Ed. by K. Hyland, M.H. Chau, and M.J.A. Handford. London: Bloomsbury, pp. 30–46.

— (2012b). *Disciplinary identifies: Individuality and community in academic discourse*. Cambridge: Cambridge University Press.

— (2013). "Researching writing". In: *Continuum companion to research methods in applied linguistics*. Ed. by B. Paltridge and A. Phakiti. London: Bloomsbury Academic, pp. 191–204.

Hyland, K. and L. Hamp-Lyons (2002). "EAP: Issues and directions". In: *Journal of English for Academic Purposes* 1.1, pp. 2–12. DOI: https://doi.org/10.1016/S1475-1585(02)00002-4.

Hyland, K., C. M. Huat, and M.J.A. Handford (2012). "Introduction". In: *Corpus applications in applied linguistics*. Ed. by K. Hyland, C.M. Huat, and M.J.A. Handford. London: Bloomsbury, pp. 3–12.

Hyland, K. and P. Tse (2005). "Hooking the reader: A corpus study of evaluative *that* in abstracts". In: *English for Specific Purposes* 24.2, pp. 123–139. DOI: https://doi.org/10.1016/j.esp.2004.02.002.

— (2007). "Is there an "academic vocabulary"?" In: *TESOL Quarterly* 41.2, pp. 235–253. DOI: https://doi.org/10.1002/j.1545-7249.2007.tb00058.x.

Hyon, S. (1996). "Genre in three traditions: Implications for ESL". In: *TESOL Quarterly* 30.4, pp. 693–722. DOI: https://doi.org/10.2307/3587930.

Jablin, F.M. and K. Krone (1984). "Characteristics of rejection letters and their effects on job applicants". In: *Written communication* 1.4, pp. 387–406. DOI: https://doi.org/10.1177/0741088384001004001.

Jamali, H.R. and M. Nikzad (2011). "Article title type and its relation with the number of downloads". In: *Scientometrics* 88, pp. 653–661. DOI: https://doi.org/10.1007/s11192-011-0412-z.

Jiang, F.K. and K. Hyland (2017). "Metadiscursive nouns: Interaction and cohesion in abstract moves". In: *English for Specific Purposes* 46, pp. 1–14. DOI: https://doi.org/10.1016/j.esp.2016.11.001.

Johns, A.M, A. Bawarshi, R.M. Coe, K. Hyland, B. Paltridge, M.J. Reiff, and C. Tardy (2006). "Crossing the boundaries of genre studies: Commentaries by experts". In: *Journal of Second Language Writing* 15.3, pp. 234–249. DOI: https://doi.org/10.1016/j.jslw.2006.09.001.

Johns, A.M., ed. (2002). *Genre in the classroom: Multiple perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Johnson, R., A. Watkinson, and M. Mabe (2018). *The STM Report: An overview of scientific and scholarly publishing*. 5th ed. The Hague, The Netherlands: International Association of Scientific, Technical and Medical Publishers. URL: https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf.

Jordan, M.P. (1991). "The linguistic genre of abstracts". In: *The 17th LACUS Forum, 1990*. Lake Bluff, IL: LACUS, pp. 507–527.

Kaltenböck, G. and B. Mehlmauer-Larcher (2005). "Computer corpora and the language classroom: On the potential and limitations of computer corpora in language teaching". In: *ReCALL: the Journal of EUROCALL* 17.1, p. 65. DOI: https://doi.org/10.1017/S0958344005000613.

Kaplan, A. (2017). *The conduct of inquiry: Methodology for behavioural science*. Abingdon, Oxon: Routledge.

Katz, S.M. (1996). "Distribution of content words and phrases in text and language modelling". In: *Natural Language Engineering* 2.1, pp. 15–59. DOI: https://doi.org/10.1017/S1351324996001246.

Kennedy, G. (1998). *An introduction to corpus linguistics*. Harlow: Addison Wesley Longman.

Kilgarriff, A. (2009). "Simple maths for keywords". In: *Proceedings of the Corpus Linguistics Conference*. Liverpool.

— (2012). "Getting to know your corpus". In: *Proceedings of the International Conference on Text, Speech and Dialogue*, pp. 3–15.

Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, and V. Suchomel (2014). "The Sketch Engine: ten years on". In: *Lexicography* 1.1, pp. 7–36. DOI: https://doi.org/10.1007/s40607-014-0009-9.

Kim, J. (2019). "Author-based analysis of conference versus journal publication in computer science". In: *Journal of the Association for Information Science and Technology* 70.1, pp. 71–82.

Kimball, R. and J. Caserta (2004). *The data warehouse ETL toolkit: Practical techniques for extracting, cleaning, conforming, and delivering data*. IN, Indianapolis: John Wiley & Sons.

Kincaid, J.P., R.P. Fishburne Jr, R.L. Rogers, and B.S. Chissom (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Tech. rep. Naval Technical Training Command Millington TN Research Branch.

Kirkpatrick, A. (2009). "English as the international language of scholarship: Implications for the dissemination of 'local' knowledge". In: *English as an international language: Perspectives and pedagogical issues*. Ed. by F. Sharifian. Bristol: Multilingual Matters, pp. 254–270.

Knapp, P. (1997). "Virtual grammar: Writing as affect/effect [Unpublished thesis]". PhD thesis. Sydney, Australia: University of Technology.

Knox, J. (Oct. 2013). *Genre vs. Text type*. sysfling Listserve.

Koester, A. (2012). "Building small specialised corpora". In: *The Routledge handbook of corpus linguistics*. Ed. by A. O'Keeffe and M.l McCarthy. Abingdon, Oxon: Routledge, pp. 66–79.

Koyamada, K., Y. Onoue, M. Kioka, T. Uetsuji, and K. Baba (2018). "Visualization of JOV abstracts". In: *Journal of Visualization* 21.2, pp. 309–319. DOI: https://doi.org/10.1007/s12650-017-0451-5.

Kress, G.R. (2003). *Literacy in the new media age*. New York, NY: Psychology Press.

Kuhn, T.S. (2012). *The structure of scientific revolutions*. 4th ed. Chicago, IL: University of Chicago.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia, PA: University of Pennsylvania Press.

Landis, J.R. and G.G. Koch (1977). "The measurement of observer agreement for categorical data". In: *Biometrics* 33 (1), pp. 159–174. DOI: https://doi.org/10.2307/2529310.

Lane, S., A. Karatsolis, and L. Bui (2015). "Graphical abstracts: A taxonomy and critique of an emerging genre". In: *Proceedings of the 33rd Annual International Conference on the Design of Communication*, pp. 1–9.

Lareo Martín, I. and A. Montoya Reyes (2007). "Scientific writing: Following Robert Boyle's principles in experimental essays - 1704 and 1998". In: *Revista alicantina de estudios ingleses* 20, pp. 119–137.

Lassen, I. (2006). "Is the press release a genre? A study of form and content". In: *Discourse Studies* 8.4, pp. 503–530. DOI: https://doi.org/10.1177/1461445606061875.

Latour, B. and S. Woolgar (1979). *Laboratory life: The social construction of scientific facts*. Beverly Hills, CA: Sage.

— (1986). *Laboratory life: The Construction of scientific facts*. 2nd Edition. Princeton, NJ: Princeton University.

Lau, H. (2004). "The structure of academic journal abstracts written by Taiwanese PhD students". In: *Taiwan Journal of TESOL* 1.1, pp. 1–25.

Lave, J. and E. Wenger (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Lee, D. (2008). "Corpora and discourse analysis". In: *Advances in discourse studies*. Ed. by V.K. Bhatia, J. Flowerdew, and R.H. Jones. London: Routledge, pp. 86–99.

— (2010). "What corpora are available". In: *The Routledge handbook of corpus linguistics*. London: Routledge, pp. 107–121.

Lee, D. and J.M. Swales (2006). "A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora". In: *English for Specific Purposes* 25.1, pp. 56–75. DOI: https://doi.org/10.1016/j.esp.2005.02.010.

Leech, G. (1991). "The state of the art in corpus linguistics". In: *English corpus linguistics*. Ed. by K. Aijmer and B. Altenberg. London: Routledge, pp. 8–30. DOI: https://doi.org/10.4324/9781315845890.

— (1992). "Corpora and theories of linguistic performance". In: *Directions in corpus linguistics*. Ed. by J. Svartvik. Berlin, Germany: Mouton de Gruyter, pp. 105–122.

— (1993). "Corpus annotation schemes". In: *Literary and Linguistic Computing* 8.4, pp. 275–281. DOI: https://doi.org/10.1093/llc/8.4.275.

— (1997). "Introducing corpus annotation". In: *Corpus annotation: Linguistic information from computer text corpora*. Ed. by R. Garside, G. Leech, and A. McEnery. London: Longman, pp. 1–18.

— (2007). "New resources, or just better old one?" In: *Corpus linguistics and the web*. Ed. by M. Hundt, N. Nesselhauf, and C. Biewer. Amsterdam: Rodopi, pp. 134–149.

Leech, G. and E. Eyes (1997). "Syntactic annotation: Treebanks". In: *Corpus annotation: Linguistic information from computer text corpora*. London: Longman, pp. 34–52.

Leonelli, S. (2007). "Arabidopsis, the botanical Drosophila: From mouse cress to model organism". In: *Endeavour* 31.1, pp. 34–38. DOI: https://doi.org/10.1016/j.endeavour.2007.01.003.

Lewin, B.A., J. Fine, and L. Young (2001). *Expository discourse: A genre-based approach to social science research texts*. New York: Continuum.

Lewis, M. (1997). "Pedagogical implications of the lexical approach". In: *Second language vocabulary acquisition: A rationale for pedagogy*. Ed. by J. Coady and T.N. Huckin. Cambridge: Cambridge University, pp. 255–270. DOI: https://doi.org/10.1017/CBO9781139524643.018.

Lieungnapar, A. and R.W. Todd (2011). "Top-down versus bottom-up approaches toward move analysis in ESP". In: *Proceedings of the International Conference on Doing Research in Applied Linguistics, King Mongkut's University of Technology Thonburi*, pp. 21–22.

Lillis, T. and M.J. Curry (2010). *Academic Writing in a global context: The politics and practices of publishing in English*. Abingdon, Oxon: Routledge.

Lock, S. (1988). "Structured abstracts". In: *British Medical Journal* 297, p. 156. DOI: https://doi.org/10.1136/bmj.297.6642.156.

Loper, E. and S. Bird (2002). "NLTK: the natural language toolkit". In: *arXiv preprint cs/0205028*.

Lorés, R. (2004). "On RA abstracts: From rhetorical structure to thematic organization". In: *English for Specific Purposes* 23 (3), pp. 280–302. DOI: https://doi.org/10.1016/j.esp.2003.06.001.

Louw, B. (1993). "Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies". In: *Text and technology: In honour of John Sinclair*. Amsterdam: John Benjamins, pp. 157–176.

Lucier, P. (2012). "The origins of pure and applied science in gilded age America". In: *ISIS* 103.3, pp. 527–536. URL: https://www.jstor.org/stable/10.1086/667976.

Lumsden, M.D. (1994). "Syntparse: A program for parsing English texts: By Hristo Georgive-Good". In: *Computers & Education* 23 (4), pp. 319–321. DOI: https://doi.org/10.1016/0360-1315(94)90020-5.

Marcus, M., B. Santorini, and M.A. Marcinkiewicz (1993). "Building a large annotated corpus of English: The Penn Treebank". In.

Markkanen, R. and H. Schröder (1992). "Hedging and its linguistic realizations in German, English and Finnish philosophical texts: A case study". In: *Fachsprachliche Miniaturen*. Frankfurt: Peter Lang, pp. 121–130.

Martin, J.R. (1992). *English Text: System and structure*. Amsterdam, Philadelphia: John Benjamins.

— (1993). "Life as a noun: Arresting the universe in science and humanities". In: *Writing science: Literacy and discursive power*. Ed. by M.A.K. Halliday and J.R. Martin. Washington, DC: The Falmer Press, pp. 221–267.

Martin, J.R. and D. Rose (2007). *Working with discourse: Meaning beyond the clause*. London: Continuum.

— (2012). "Genres and texts: Living in the real world". In: *Indonesian Journal of Systemic Functional Linguistics* 1.1, pp. 1–21. URL: http://alsfal2013.weebly.com/uploads/1/6/5/5/16553900/genres_and_texts.pdf.

Martıín, P.M. (2003). "A genre analysis of English and Spanish research paper abstracts in experimental social sciences". In: *English for Specific Purposes* 22.1, pp. 25–43. DOI: https://doi.org/10.1016/S0889-4906(01)00033-3.

Martıín, P.M. and S. Burgess (2006). "Reader and writer responsibility in abstracts in Spanish social sciences journals". In: *Explorations in specialized genres*. Vol. 35, pp. 43–57.

Mathet, Y., A. Widlöcher, K. Fort, C. François, O. Galibert, C. Grouin, J. Kahn, S. Rosset, and P. Zweigenbaum (2012). "Manual corpus annotation: Giving meaning to the evaluation metrics". In: *Proceedings of COLING 2012*, pp. 809–818. URL: https://www.aclweb.org/anthology/C12-2079.pdf.

Matthiessen, C.M.I.M. and M.A.K. Halliday (2009). *Systemic functional grammar: A first step into the theory*. Beijing, China: Higher Education Publishing House.

McDonald, R.J., H.J. Cloft, and D.F. Kallmes (2007). "Fate of submitted manuscripts rejected from the American Journal of Neuroradiology: Outcomes and commentary". In: *American Journal of Neuroradiology* 28.8, pp. 1430–1434. DOI: https://doi.org/10.3174/ajnr.a0766.

McEnery, T. and A. Hardie (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University.

McEnery, T. and A. Wilson (2001). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T., R. Xiao, and Y. Tono (2006). *Corpus-based language studies: An advanced resource book*. New York, NY: Taylor & Francis.

Meinke, D.W., J.M.l Cherry, C. Dean, S.D. Rounsley, and M. Koornneef (1998). "Arabidopsis thaliana: A model plant for genome analysis". In: *Science* 282.5389, pp. 662–682. DOI: https://doi.org/10.1126/science.282.5389.662.

Melander, B., J.M. Swales, and K.M. Fredickson (1997). "Journal abstracts from three academic fields in the United States and Sweden: National or disciplinary proclivities?" In: *Intellectual styles and cross-cultural communication*. Ed. by A. Duszak. Berlin: Mouton De Gruyter, pp. 251–272.

Miller, C.R. (1994). "The cultural basis of genre". In: *Genre and the new rhetoric*. Ed. by A. Feedman and P. Medway. London: Taylor and Francis, pp. 67–78.

Mindt, D. (2002). "What is a grammatical rule?" In: *From the COLT's mouth... and others': Language corpora studies, In Honour of Anna-Brita Stenström*. Ed. by L.E. Breivik and A. Hasselgren. Amsterdam, Netherlands: Brill Rodopi, pp. 197–212.

Miniwatts Marketing Group (2018). *World internet users statistics and 2015 world population stats*. URL: https://www.internetworldstats.com/stats.htm.

Mizumoto, A., S. Hamatani, and Y. Imao (2017). "Applying the bundle–move connection approach to the development of an online writing support tool for research articles". In: *Language Learning* 67.4, pp. 885–921. DOI: https://doi.org/10.1111/lang.12250.

Montgomery, S.L. (2013). *Does science need a global language?: English and the future of research*. Chicago, IL: University of Chicago Press.

Morley, J. (2004). "An initiative to develop Academic Phrasebank: A university-wide online writing resource". In: *Snaphots of innovation*. Manchester: University of Manchester.

— (2020). *Academic Phrasebank [online resource]*. URL: http://www.phrasebank.manchester.ac.uk/.

Mu, E. and M. Pereyra-Rojas (2017). *Practical decision making: An introduction to the Analytic hierarchy Process (AHP) using super decisions V2*. Cham, Switzerland: Springer, pp. 7–22.

Murphy, R. (2012). *English grammar in use*. Cambridge: Cambridge University Press.

Myers, G. (1990). "The rhetoric of irony in academic writing". In: *Written Communication* 7.4, pp. 419–455. DOI: https://doi.org/10.1177/0741088390007004001.

— (1991). *Writing biology: Texts in the social construction of scientific knowledge*. Winconsin, WC: University of Winconsin.

National Science Foundation (2018). *Federal R&D Funding, by budget function: Fiscal years 2016–18. Detailed statistical tables NSF*. Alexandria, VA.: National Center for Science and Engineering Statistics, pp. 18–308. URL: https://www.nsf.gov/statistics/2018/nsf18308/.

Nattinger, J.R and J.S. DeCarrico (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nelson, M. (2010). "Building a written corpus: What are the basics?" In: *The Routledge handbook of corpus linguistics*. Ed. by A. O'Keeffe and M. McCarthy. Oxon: Routledge, pp. 53–65.

Nwogu, K.N. (1990). *Discourse variation in medical texts: Schema, theme and cohesion on professional and journalistic accounts*. Tech. rep. Nottingham: Department of English Studies. University of Nottingham.

O'Donnell, M. (Oct. 2013). *Genre vs. Text type*. sysfling Listserve.

O'Keeffe, A. (2007). "The pragmatics of corpus linguistics". In: *Keynote paper presented at the 4th Corpus Linguistics Conference*. Birmingham: University of Birmingham.

O'Keeffe, A. and M. McCarthy, eds. (2010). *The Routledge Handbook of Corpus Linguistics*. London: Routledge.

O'Keeffe, A., M. McCarthy, and R. Carter (2007). *From Corpus to classroom*. Cambridge: Cambridge University Press.

O'Donnell, M. (2008). "The UAM CorpusTool: Software for corpus annotation and exploration". In: *Proceedings of the XXVI Congreso de AESLA*. Vol. 3. 5. Almeria, Spain.

Oakey, D. (Feb. 10, 2009). *The lexical bundle revisited: Isolexical and isotextual comparisons. English Language Research seminar*. Tech. rep. Corpus Linguistics and Discourse. University of Birmingham.

Okamura, A. (2006). "Two types of strategies used by Japanese scientists, when writing research articles in English". In: *System* 34, pp. 68–79. DOI: https://doi.org/10.1016/j.system.2005.03.006.

Okulicz-Kozaryn, A. (2013). "Cluttered writing: Adjectives and adverbs in academia". In: *Scientometrics* 96, pp. 679–681. DOI: https://www.doi.org/10.1007/s11192-012-0937-9.

Oms, S. and E. Zardini (2019). *The Sorites paradox*. Cambridge: Cambridge University Press. DOI: 10.1017/9781316683064.

Osborne, J. (2004). "Top-down and bottom-up approaches to corpora in language teaching". In: *Applied corpus linguistics*. Leiden, Netherlands: Brill Rodopi, pp. 251–265.

Paice, C.D. (1980). "The automatic generation of literature abstracts: An approach based on the identification of self-indicating phrases". In: *Proceedings of the 3rd*

annual ACM Conference on Research and Development in Information Retrieval. Ed. by R.N. Oddy, C.J. Rijsbergen, and P.W. Williams. Butterworth & Co., pp. 172–191.

Paltridge, B. (2006). *Discourse analysis: An introduction*. London: Continuum.

Parkhurst, C. (1990). "The composition process of science writers". In: *English for Specific Purposes* 9.2, pp. 169–179. DOI: https://doi.org/10.1016/0889-4906(90)90006-X.

Passonneau, R. (2006). "Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation". In: *Fifth International Conference on Language Resources and Evaluation*, pp. 831–836.

Pawley, A. and F.H. Syder (1983). "Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar". In: *Journal of Pragmatics* 7.5, pp. 551–579. DOI: https://doi.org/10.1016/0378-2166(83)90081-4.

Pellegrino, E.D. (1964). "Patient care–Mystical research or researchable mystique?" In: *Clinical Research* 12.4, pp. 421–425.

Pendar, N. and E. Cotos (2008). "Automatic identification of discourse moves in scientific article introductions". In: *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 62–70. URL: https://www.aclweb.org/anthology/W08-0908.pdf.

Peshkin, A. (1988). "In search of subjectivity. One's own". In: *Educational Researcher* 17.7, pp. 17–21. DOI: https://doi.org/10.3102/0013189X017007017.

Phillips, M. (1989). *Lexical structure of text*. Birmingham: University of Birmingham.

Pho, P. D. (2008). "Research article abstracts in applied linguistics and educational technology: A study of linguistic realizations of rhetorical structure and authorial stance". In: *Discourse studies* 10.2, pp. 231–250. DOI: https://www.doi.org/10.1177/1461445607087010.

Pillai, A.B. (2017). *Software architecture with Python*. Packt Publishing Ltd.

Plavén-Sigray, P., G.J. Matheson, B. C. Schiffler, and W.H. Thompson (2017). "The readability of scientific texts is decreasing over time". In: *Elife* 6, e27725.

Potter, J.E.R. and G. Talukder (2003). "Past versus present: The importance of tense in patent application examples". In: *Nature Biotechnology* 21.11, pp. 1397–1398. DOI: https://doi.org/10.1038/nbt1103-1397.

Pustejovsky, J., J.M. Castano, R. Ingria, R.r Sauri, R.J. Gaizauskas, A. Setzer, G. Katz, and D.R. Radev (2003). "TimeML: Robust specification of event and temporal expressions in text". In: *New Directions in Question Answering* 3, pp. 28–34.

Quirk, R. and S. Greenbaum (1993). *A university grammar of English*. Hong Kong: Longman.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Raghavan, P., E. Fosler-Lussier, and A.M. Lai (2012). "Inter-annotator reliability of medical events, coreferences and temporal relations in clinical narratives by annotators with varying levels of clinical expertise". In: *Proceedings of 2012 AMIA*

Annual Symposium. American Medical Informatics Association, pp. 1366–1374. URL: http://www.ncbi.nlm.nih.gov/pubmed/23304416.

Rasmussen, C.E. and Z. Ghahramani (2001). "Occam's razor". In: *Advances in neural information processing systems*, pp. 294–300.

Rau, G. (2019). *Writing for engineering and science students: Staking your claim*. London: Taylor & Francis.

Rau, G. and A. Antink-Meyer (2020). "Distinguishing science, engineering, and technology". In: *Nature of science in science instruction*. Ed. by W. McComas. Cham, Switzerland: Springer, pp. 159–176. DOI: https://doi.org/10.1007/978-3-030-57239-6_8.

Ray, J., M. Berkwits, and F. Davidoff (2000). "The fate of manuscripts rejected by a general medical journal". In: *The American Journal of Medicine* 109.2, pp. 131–135. DOI: https://doi.org/10.1016/S0002-9343(00)00450-2.

Reich, E.S. (2013). "Science publishing: The golden club". In: *Nature News* 502.7471, p. 291.

Ren, H. and Y. Li (2011). "A comparison study on the rhetorical moves of abstracts in published research articles and master's foreign-language theses". In: *English Language Teaching* 4.1, pp. 162–166. DOI: https://doi.org/10.5539/elt.v4n1p162.

Reppen, R. (2012). "Building a corpus: What are the key considerations?" In: *The Routledge handbook of corpus linguistics*. Ed. by A. O'Keeffe and M. McCarthy. Abingdon, Oxon: Routledge, pp. 81–93.

Resnik, D.B (2000). "A pragmatic approach to the demarcation problem". In: *Studies in History and Philosophy of Science Part A* 31.2, pp. 249–267. DOI: https://doi.org/10.1016/S0039-3681(00)00004-2.

Roberts, A., R.t Gaizauskas, M. Hepple, N. Davis, G. Demetriou, Y. Guo, J.S. Kola, I. Roberts, A. Setzer, A. Tapuria, et al. (2007). "The CLEF corpus: Semantic annotation of clinical text". In: *Proceedings of the AMIA Annual Symposium*. American Medical Informatics Association, pp. 625–629.

Rodrigues, N., L.G. Magalhães, J. Moura, and A. Chalmers (2008). "Automatic Reconstruction of Virtual Heritage Sites". In: *The 9th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST)*. Ed. by M. Ashley, S. Hermon, A. Proenca, and K. Rodriguez-Echavarria, pp. 39–46.

Rose, D. (Oct. 2013a). *Genre vs. Text type*. sysfling Listserve.

— (2013b). "The Routledge handbook of discourse analysis". In: ed. by J.P. Gee and M. Handford. London: Routledge, pp. 209–225.

Rounds, P. (1982). *Hedging in written academic discourse: Precision and flexibility*. Ann Arbor, MI: University of Michigan.

Sagae, K. and A. Lavie (2003). "Combining rule-based and data-driven techniques for grammatical relation extraction in spoken language". In: *Proceedings of the Eighth International Conference on Parsing Technologies*. URL: https://aclanthology.org/W03-3019.pdf.

Sagi, I. and E. Yechiam (2008). "Amusing titles in scientific journals and article citation". In: *Journal of Information Science* 34 (5), pp. 680–687. DOI: https://doi.org/10.1177/0165551507086261.

Salager-Meyer, F. (1990). "Discoursal flaws in medical English abstracts: A genre analysis per research- and text-type". In: *Text* 10.4, pp. 365–84. DOI: https://doi.org/10.1515/text.1.1990.10.4.365.

— (1992). "A text-type and move analysis study of verb tense and modality distribution in medical English abstracts". In: *English for Specific Purposes* 11.2, pp. 93–113. DOI: https://doi.org/10.1016/S0889-4906(05)80002-X.

— (1994). "Hedges and textual communicative function in medical English written discourse". In: *English for Specific Purposes* 13 (2), pp. 149–171. DOI: https://doi.org/10.1016/0889-4906(94)90013-2.

Salager-Meyer, F., A.A. Ariza, and N. Zambrano (2003). "Scimitar, the dagger and the glove: Intercultural differences in the rhetoric of criticism in Spanish, French and English Medical Discourse (1930–1995)". In: *English for Specific Purposes* 22.3, pp. 223–247. DOI: https://doi.org/10.1016/S0889-4906(02)00019-4.

Samar, R.G., H. Talebzadeh, G.R. Kiany, and R. Akbari (2014). "Moves and steps to sell a paper: A cross-cultural genre analysis of applied linguistics conference abstracts". In: *Text & Talk* 34.6, pp. 759–785. DOI: https://doi.org/10.1515/text-2014-0023.

Sampson, G. and D. McCarthy, eds. (2005). *Corpus linguistics: Readings in a widening discipline*. London: Continuum.

Samraj, B. (2002). "Introductions in research articles: Variations across disciplines". In: *English for Specific Purposes* 21, pp. 1–17. DOI: https://doi.org/10.1016/S0889-4906(00)00023-5.

— (2005). "An exploration of genre set: Research article abstracts and introduction in two disciplines". In: *English for Specific Purposes* 24 (2), pp. 141–156. DOI: https://doi.org/10.1016/j.esp.2002.10.001.

Sarkar, S. (1992). "Models of reduction and categories of reductionism". In: *Synthese* 91.3, pp. 167–194. DOI: https://www.doi.org/10.1007/BF00413566.

Saussure, F. (1966). *Course in general linguistics*. Ed. by Charles Bally and Albert Sechehaye. Trans. by Wade Baskin. New York, NY: McGraw-Hill.

Schickore, J. (2008). "Doing science, writing science". In: *Philosophy of Science* 75.3, pp. 323–343. DOI: https://doi.org/10.1086/592951.

Schiffrin, D. (1994). *Approaches to discourse*. Oxford: Blackwell Publishers Ltd.

Schmidt, R. (2012). "Attention, awareness, and individual differences in language learning". In: *Perspectives on Individual Characteristics and Foreign Language Education* 6, p. 27.

Schütze, C.T. and J. Sprouse (2013). "Judgment data". In: *Research methods in linguistics*. Ed. by R.J. Podesva and D. Sharma. Cambridge: Cambridge University Press, pp. 27–50.

Scott, M. (1997). "PC analysis of key words – and key key words". In: *System* 25.1, pp. 1–13. DOI: https://doi.org/10.1016/S0346-251X(97)00011-0.

— (2009). "In search of a bad reference corpus". In: *What's in a word-list? Investigating word frequency and keyword extraction*. Ed. by D.E. Archer. London: Taylor and Francis, pp. 79–92. DOI: https://doi.org/10.4324/9781315547411.

— (2019). *WordSmith Tools (version 8.0)*. URL: https://lexically.net/wordsmith/index.html.

Scott, M. and C. Tribble (2015). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam/Philadelphia, PA: John Benjamins Publishing.

Sebastiani, F. (2002). "Machine learning in automated text categorization". In: *ACM Computing*, pp. 1–47. DOI: https://doi.org/10.1145/505282.505283.

Seliger, H.W. and E. Shohamy (1989). *Second language research methods*. Oxford: Oxford University Press.

Selinker, L. (1979). "On the use of informants in discourse analysis and 'language for specialized purposes'". In: *International Review of Applied Linguistics in Language Teaching* 17.1-4, pp. 189–216. DOI: https://doi.org/10.1515/iral.1979.17.1-4.189.

Sharma, S. and J.E. Harrison (2006). "Structured abstracts: Do they improve the quality of information in abstracts?" In: *American Journal of Orthodontics and Dentofacial Orthopedics* 130.4, pp. 523–530. DOI: https://doi.org/10.1016/j.ajodo.2005.10.023.

Silva, T. (1993). "Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications". In: *TESOL Quarterly* 27.4, pp. 665–677. DOI: https://doi.org/10.2307/3587400.

Simionescu, M. and E. Simion (2004). "Scientific lingua franca and National Languages at the Crossroads". In: *ALLEA biennial yearbook. Critical topics in science and scholarship*. Ed. by P. Drenth and J. Schroots. Amsterdam: ALLEA, pp. 129–133. URL: https://allea.org/wp-content/uploads/2016/02/Simionescu_-Lingua_Franca.pdf.

Sinclair, J.M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

— (2001). "Preface". In: *Small corpus studies and ELT: Theory and practice*. Ed. by M. Ghadessy, A. Henry, and R.L. Roseberry. John Benjamins Publishing.

— (2004a). "Corpus and text — Basic principles". In: *Developing linguistic corpora: A guide to good practice*. Ed. by M. Wynne. Oxford: Oxbow Books.

— (2004b). "The search for units of meaning". In: *Textus* 9 (1), pp. 75–106.

— (2004c). *Trust the text: Language corpus and discourse*. London: Routledge.

— (2008). "Borrowed ideas". In: *Language, people, numbers*. Ed. by O Gerbig A.and Mason. Amsterdam: Rodopi BV, pp. 21–42.

Sinclair, J.M. and M. Coulthard (1975). *Towards an analysis of discourse: The English used by teachers and pupils*. Oxford: Oxford University Press.

Snow, C.P. (1963). *The two cultures*. Cambridge: Cambridge University Press.

Soler, V. (2011). "Comparative and contrastive observations on scientific titles written in English and Spanish". In: *English for Specific Purposes* 30 (2), pp. 30–124. DOI: https://doi.org/10.1016/j.esp.2010.09.002.

Sollaci, L.B. and M.G. Pereira (2004). "The introduction, methods, results, and discussion (IMRAD) structure: A fifty-year survey". In: *Journal of the Medical Library Association* 92.3, p. 364.

Stefanowitsch, A. and S.Th. Gries (2003). "Collostructions: Investigating the interaction of words and constructions". In: *International Journal of Corpus Linguistics* 8.2, pp. 209–243. DOI: https://doi.org/10.1075/ijcl.8.2.03ste.

Stiny, G. and W.J. Mitchell (1978). "The Palladian grammar". In: *Environment and planning B: Planning and design* 5.1, pp. 5–18. DOI: https://doi.org/10.1068/b050005.

Stotesbury, H. (2003). "Evaluation in research article abstracts in the narrative and hard sciences". In: *Journal of English for Academic Purposes* 2.3, pp. 327–341. DOI: https://doi.org/10.1016/S1475-1585(03)00049-3.

Strauss, A. and J.M. Corbin (1997). *Grounded theory in practice*. London: Sage.

Stubbs, M. (1995). "Collocations and semantic profiles: On the cause of the trouble with quantitative studies". In: *Functions of Language* 2.1, pp. 23–55. DOI: https://doi.org/10.1075/fol.2.1.03stu.

— (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Oxford: Blackwell.

— (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell Publishers.

— (2007). "On texts, corpora and models of language". In: *Text, discourse and corpora*. Ed. by M. Hoey, M. Mahlberg, M. Stubbs, and W. Teubert. London: Continuum, pp. 163–190.

Suntara, W. (2013). "Rhetorical variations in research article abstracts and introductions in linguistics and applied linguistics". PhD thesis. Suranaree University of Technology.

Suntara, W. and S. Usaha (2013). "Research article abstracts in two related disciplines: Rhetorical variation between linguistics and applied linguistics". In: *English Language Teaching* 6.2, pp. 84–99. DOI: https://doi.org/10.5539/elt.v6n2p84.

Swales, J.M. (1988). "Discourse communities, genres and English as an international language". In: *World Englishes* 7.2, pp. 211–220. DOI: https://doi.org/10.1111/j.1467-971X.1988.tb00232.x.

— (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

— (1996). "Occluded genres in the academy: The case of the submission letter". In: *Academic writing: Intercultural and textual issues*. Ed. by E. Ventola and A. Mauranen. Amsterdam, The Netherlands: John Benjamins, pp. 45–58.

— (1997). "English as *Tyrannosaurus rex*". In: *World Englishes* 16.3, pp. 373–382. DOI: https://doi.org/10.1111/1467-971X.00071.

Swales, J.M. (1998). *Other floors, other voices: A textography of a small university building*. Routledge.

— (2004). *Research genres: Explorations and applications*. Cambridge: Cambridge University Press.

Swales, J.M. and C.B. Feak (2009). *Abstracts and the writing of abstracts*. Ann Arbor, MI: Michigan University Press.

Teubert, W. and R. Krishnamurthy (2007). "General introduction". In: *Corpus linguistics: Critical concepts in linguistics*. Ed. by W. Teubert and R. Krishnamurthy. London: Routledge, pp. 1–37.

Thompson, G. and S. Hunston (2006). *System and corpus: Exploring connections*. London: Equinox.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.

— (2010). "Theoretical overview of the evolution of corpus linguistics". In: *The Routledge handbook of corpus linguistics*. Ed. by A. O'Keeffe and M. McCarthy. Oxon: Routledge, pp. 14–18.

Tort, A.B.L., Z.H. Targino, and O.B. Amaral (2012). "Rising publication delays inflate journal impact factors". In: *PLoS One* 7.12, e53374. DOI: https://doi.org/10.1371/journal.pone.0053374.

Tribble, C. (1997). "Improvising corpora for ELT: Quick-and-dirty ways of developing corpora for language teaching". In: University of Lodz, Poland. URL: https://www.ctribble.co.uk/text/Tribble_C_Palc_97.pdf.

Tsamir, P. and D. Tirosh (2002). "Intuitive beliefs, formal definitions and undefined operations: Cases of division by zero". In: *Beliefs: A hidden variable in mathematics education?* Ed. by G.C. Leder, E. Pehkonen, and G. Törner. Dordrecht, Netherlands: Springer, pp. 331–344. DOI: https://doi.org/10.1007/0-306-47958-3_19.

Tseng, F.P. (2011). "Analyses of move structure and verb tense of research article abstracts in applied linguistics journals". In: *International Journal of English linguistics* 1.2, p. 27. DOI: https://doi.org/10.5539/ijel.v1n2p27.

Tu, P. and S. Wang (2013). "Corpus-based research on tense analysis and rhetorical structure in journal article abstracts". In: *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, pp. 102–107. URL: https://www.aclweb.org/anthology/Y13-1008.pdf.

UNESCO (2018). *United Nations Educational, Scientific and Cultural Organisation Science Report: Towards 2010, executive summary*. URL: https://unesdoc.unesco.org/ark:/48223/pf0000235407.

Upton, T.A. and M.A. Cohen (2009). "An approach to corpus-based discourse analysis: The move analysis as example". In: *Discourse Studies* 11.5, pp. 585–605.

Van Dijk, T.A. (1980). *Macrostructures*. Hillsdale, NJ: Laurence Earlbaum.

Vande Koppel, W.J. (1994). "Some characteristics and functions of grammatical subjects in scientific discourse". In: *Written Communication* 11.4, pp. 534–564. DOI: https://doi.org/10.1177/0741088394011004004.

Váradi, T. (2001). "The linguistic relevance of corpus linguistics". In: *Proceedings of the Corpus Linguistics 2001 Conference. University Centre for Computer Corpus Research on Language Technical Papers*. Ed. by P. Rayson, A. Wilson, T. McEnery, A. Hardie, and S. Khoja. Vol. 13. Lancaster University, pp. 587–593.

Vassileva, I. (2002). "Speaker-audience interaction: The case of Bulgarians presenting in English". In: ed. by E. Ventola, C. Shalom, and S. Thompson. Frankfurt am Main: Peter Lang, pp. 255–276.

Venable, J. (2006). "The role of theory and theorising in design science research". In: *Proceedings of the 1st International Conference on Design Science in Information Systems and Technology (DESRIST 2006)*, pp. 1–18.

Ventola, E. (1992). "Writing scientific English: Overcoming intercultural problems". In: *International Journal of Applied Linguistics* 2.2, pp. 191–220. DOI: https://doi.org/10.1111/j.1473-4192.1992.tb00033.x.

— (1994). "Abstracts as an object of linguistic study". In: ed. by S. Čmejrková, F. Daneš, and E. Havlov. Tübingen: Gunter Narr, pp. 333–352.

Virtanen, T. (2009). "Discourse linguistics meets corpus linguistics: Theoretical and methodological issues in the troubled relationship". In: *Corpus linguistics: Refinements and reassessments*. Vol. 69. Language and Computers, pp. 49–65. DOI: https://doi.org/10.1163/9789042025981_005.

Volodina, E., I. Pilán, S. R. Eide, and H. Heidarsson (2014). "You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a second language". In: *Proceedings of the Third Workshop on NLP for Computer-Assisted Language Learning*, pp. 128–144.

Wallis, S.A. (2003). "Completing parsed corpora". In: *Treebanks*. Ed. by A. Abeillé. New York, NY: Springer Science+Business Media, pp. 61–71.

— (2007). "Annotation, retrieval and experimentation". In: *Studies in Variation, Contacts and Change in English* 1. URL: http://www.helsinki.fi/varieng/series/volumes/01/wallis/.

Wallis, S.A. and G. Nelson (2001). "Knowledge discovery in grammatically analysed corpora". In: *Data Mining and Knowledge Discovery* 5, pp. 307–340. DOI: https://doi.org/10.1023/A:1011453128373.

Wang, Y. and Y. Bai (2007). "A corpus-based syntactic study of medical research article titles". In: *System* 35 (3), pp. 388–399. DOI: https://doi.org/10.1016/j.system.2007.01.005.

Wickham, H. (2014). "Tidy data". In: *Journal of Statistical Software* 59 (10), pp. 1–23. URL: https://www.jstatsoft.org/issue/view/v059.

Willis, D. (2003). *Rules, patterns and words: Grammar and lexis in English language teaching*. Cambridge: Cambridge University Press.

Wray, A. (2005). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Yakhontova, T. (1999). "Cultural variation in the genre of conference abstract: Rhetorical and linguistic dimensions". In: *Symposium on English as a Conference Language in Europe*. Wittenberg, Germany, pp. 14–17.

— (2019). "" The authors have wasted their time...": Genre features and language of anonymous peer reviews." In: *Topics in Linguistics* 20.2, pp. 67–89. DOI: https://doi.org/10.2478/topling-2019-0010.

Yanchun, L (2007). "A genre analysis of English and Chinese research article abstracts published in linguistic and mathematic journals". MA thesis. Chongquing, China.

Yang, J.T. (1995). *An outline of scientific writing: For researchers with English as a foreign language*. Singapore: World Scientific Publishing Company.

Yang, R. and D. Allison (2003). "Research articles in applied linguistics: Moving from results to conclusions". In: *English for Specific Purposes* 22.4, pp. 365–385. DOI: https://doi.org/10.1016/S0889-4906(02)00026-1.

Yimam, S.M., I. Gurevych, R. Eckart de Castilho, and C. Biemann (2013). "WebAnno: A flexible, web-based and visually supportedsystem for distributed annotations". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 1–6. URL: https://www.aclweb.org/anthology/P13-4001.

Yoon, J. and E. Chung (2017). "An investigation on graphical abstracts use in scholarly articles". In: *International Journal of Information Management* 37.1, pp. 1371–1379. DOI: https://doi.org/10.1016/j.ijinfomgt.2016.09.005.

Yule, G. (1998). *Explaining English grammar: A guide to explaining grammar for teachers of English as a second or foreign language.* Oxford: Oxford University Press.

Zeldes, A. (2018). *Multilayer corpus studies*. Oxon: Routledge.