Lexico-grammatical properties of abstracts and research articles

A corpus-based study of scientific discourse from multiple disciplines

GENEHMIGTE DISSERTATION

zur Erlangung eines Grades des Doktors der Philosophie
im Fachbereich Gesellschafts- und Geschichtswissenschaften
an der Technischen Universität Darmstadt

Referentinnen:
Prof. Dr. Elke Teich
Prof. Dr. Nina Janich

vorgelegt von
Mônica Holtz, M.A.
aus Rio de Janeiro, Brasilien

Tag der Einreichung: 10.01.2011
Tag der mündlichen Prüfung: 27.05.2011

D17

Darmstadt
2011

Abstract

Research articles are acknowledged to be the most important form of scientific discourse. Abstracts are, apart from the title, the first meeting of readers with research articles. Independently of their traditional purpose to summarize research articles, abstracts have become crucial for readers in the decision process of reading the text further, especially nowadays due to the vast amount of scientific publications. The growing importance of abstracts in academia and the few existing research focused on these have motivated this present research, which explored the relationship between these two text types in a broader linguistic context and investigated the linguistic differences between abstracts and research articles based on the quantitative analysis of the distribution of selected features.

This research is rooted in Systemic Functional Linguistics, a sophisticated linguistic model, making possible the analysis of the relations between language and different social contexts and allowing a detailed investigation of discourse variation based on the analysis of linguistic features. This theory suggests a corpus linguistic methodology and the interest in functional variation of language is inherent in it. For this study, a corpus of English abstracts and research articles of the disciplines of computer science, linguistics, biology, and mechanical engineering was compiled and processed according to current practices in corpus linguistics.

The study applied a twofold methodology. First, a deductive empirical analysis was performed, by which selected features were quantitatively determined and statistically evaluated for significance and hypothesis testing. Then, an inductive empirical analysis was conducted that corroborated the results of the deductive analysis that ascertained the adequacy of the hypotheses and features chosen. The results indicated that abstracts and their research articles are significantly distinct from each other together with a clear domain specific variation.

This research contributes further to the linguistic investigation of scientific discourse. Not only linguists interested in language variation profit from the results acquired here. Such a research can contribute to the area of English for Special Purposes, having pedagogical applications in teaching of contemporary academic and research English inasmuch as understanding a certain discipline and practices of its community involves understanding their literacy.

Zusammenfassung

Wissenschaftliche Aufsätze sind als wichtigste Form des wissenschaftlichen Diskurses anerkannt. Abstracts sind, neben dem Titel, die erste Begegnung der Leserschaft mit dem wissenschaftlichen Aufsatz. Unabhängig von ihrer traditionellen Funktion, den wissenschaftlichen Aufsatz zusammenzufassen, sind Abstracts fundamental für den Entscheidungsprozess der Leserschaft den Text weiterzulesen, insbesondere heutzutage, bedingt durch die große Anzahl an wissenschaftlichen Veröffentlichungen. Die wachsende Bedeutung von Abstracts im wissenschaftlichen Diskurs und die wenigen wissenschaftlichen Studien, die sich mit ihnen beschäftigen, waren Motivation für diese Arbeit, in der das Verhältnis zwischen diesen beiden Textsorten, in einem breiteren linguistischen Kontext, und die linguistischen Unterschiede zwischen Abstracts und wissenschaftlichen Aufsätzen, basierend auf einer quantitativer Analyse der Verteilung ausgewählter Merkmale, untersucht wurden.

Diese Arbeit basiert auf der Systemisch Funktionalen Linguistik, einem komplexen linguistischen Modell, welches die Analyse der Relationen zwischen Sprache und verschiedenen sozialen Kontexten, wie auch die detaillierte Untersuchung von Diskursvariationen, basierend auf der Analyse linguistischer Merkmale, ermöglicht. Diese Theorie legt eine korpuslinguistische Methodologie nahe und hat ein inhärentes Interesse an der funktionalen Variation von Sprache. Für diese Studie wurde ein Korpus aus englischen Abstracts und wissenschaftlichen Aufsätzen aus den Disziplinen Informatik, Linguistik, Biologie und Maschinenbau, entsprechend den aktuell gültigen Methoden der Korpuslinguistik, erstellt und prozessiert.

In dieser Arbeit wurde ein zweifältiger methodologischer Ansatz gewählt. Zuerst wurde eine deduktive empirische Analyse durchgeführt, durch die ausgewählte Merkmale quantitativ bestimmt, und statistisch bezüglich der Signifikanz und zur Prüfung der Hypothesen bewertet wurden. Dann wurde eine induktive empirische Analyse durchgeführt, um die Ergebnisse der deduktiven Analyse zu erhärten, und um die Adäquanz der Hypothesen und der gewählten Merkmale zu bestätigen. Die Ergebnisse zeigten, dass sich Abstracts und die dazugehörigen wissenschaftlichen Aufsätze signifikant voneinander unterscheiden, mit klarer domänenspezifischer Variation.

Diese Dissertation leistet weiterhin einen Beitrag zur linguistischen Untersuchung der Wissenschaftssprache. Nicht nur Linguisten, die an Sprachvariationen interessiert sind, profitieren von den hier erlangten Ergebnissen. Diese Arbeit kann auch einen Beitrag auf dem Gebiet des *English for Special Purposes* leisten, als hier pädagogi-

sche Anwendungen für die Lehre von zeitgenössischem Akademischen Englisch abgeleitet werden können, insofern, als das Verstehen einer bestimmten Disziplin, und der Gepflogenheiten ihrer Wissenschaftsgemeinde, das Verständnis ihrer Literalität voraussetzt.

I dedicate this thesis to my husband, Frithjof,
who has always helped me and believed that I could do it.

Acknowledgements

I wish to thank, first and foremost, my supervisor Prof. Dr. Elke Teich for her continuous support, guidance and encouragement over the last few years. Thank you for the countless discussions we have had about this research over coffee and cigarettes. They have been tremendously valuable for me.

I would like to thank Prof. Dr. Nina Janich, Prof. Dr. Julika Griem and Prof. Dr. Dieter Schott for acting as members of the examination committee and for their constructive questions, suggestions and comments about this research.

I am deeply grateful to my colleague and friend Dr. Sabine Bartsch, who has patiently listened to all my problems and helped me over the last (many) years. I am also indebted to Richard Eckart de Castilho for the development of several computational linguistic tools and for helping me retrieving information from the data. Thank you for all the vivid discussions we have had!

Special thanks go to several people who have contributed with comments and suggestions at different stages of this research: Stefania Degaetano, Prof. Dr. Stefan Evert, Dr. Peter Fankhauser, Ngozi Gamisch, Tatsiana Marković, Prof. Dr. Stella Neumann and Anke Schulz.

I would not have been able to complete this work without the support of my family. Many thanks go to my parents, my sisters, especially Alessandra, my brother and my mother-in-law, who, even not knowing anything about this research, have motivated me endlessly.

I am grateful for the financial support from the *Technische Universität Darmstadt* and from the German Research Foundation (DFG - *Deutsche Forschungsgemeinschaft*, grant TE 198/1-1).

All remaining errors are, of course, entirely mine.

Contents

List of Abbreviations

AbstRA	Abstracts and Research Articles Corpus
ANOVA	Analysis of variance
cf.	*confer*
CL	Corpus linguistics
DaSciTex	Darmstadt Scientific Text Corpus
e.g.	*exempli gratia*
ESP	English for Special Purposes
i.e.	*id est*
p.	page
PCA	Principal component analysis
PoS	Part-of-speech
RA	Research article
SFL	Systemic Functional Linguistics
STTR	Standardized type/token ratio
TTR	Type/token ratio

List of Figures

List of Tables

Chapter 1

Introduction

The classic stereotype of scientists as individuals working alone in lab coats or introspectively reflecting about phenomena in the world does not portray the contemporary practices of the scientific community. In fact, scientists spend a considerable amount of their time on the elaboration, analysis, presentation and exchange of knowledge. This knowledge is construed and expressed through language, more specifically, scientific discourse, a functional variation of language with its own technical terminology and grammar (Halliday et al. 1964; Halliday & Martin 1993; Martin 1992b; Martin & Veel 1998). Scientists write as members of a group adopting practices of discourse and complying with their own understanding and perception of the world. In order to engage with such a community, one must be able to use its language accordingly. Scientific discourse is, therefore, a valuable source of information about social semiotic interactions within the scientific community.

Scientific discourse has been the subject of quite a few linguistic studies. Linguistic research on this topic ranges from the description of "scientific writing" (e.g., Banks 2008; Halliday & Martin 1993; Ventola 1996) up to analyses of specific discourse fields (e.g., O'Halloran 2005 on mathematics) and genres (e.g., Swales 1981, 1990, 2004; Ventola 1997). Halliday & Martin (1993: 8) argue that "scientific language just foregrounds the constructive potential of language as a whole". Therefore, research on scientific discourse is relevant not only for the characterization of this variation in particular, but more widely, for language as such. Additionally, language variation spread gradually to other discourses rather than science, thereby influencing the general interpretation of human experience.

> Every text, from the discourses of technocracy and bureaucracy to the television magazine and the blurb on the back of the cereal packet, is in some way affected by the modes of meaning that evolved as the scaffolding for scientific knowledge. In other words, the language of science has become the language of literacy.
>
> (Halliday & Martin 1993: 11)

Members of scientific communities traditionally publicize their knowledge mainly in the form of books, monographs, theses, dissertations, presentations, and research articles (RA). From all possibilities of realizing scientific discourse, "the grand master narrative of modernism" (Montgomery 1996: 2), RA became the most important one. The value attached to it increased considerably since the publication of the first scientific journal in Europe, the *Le Journal des sçavans* and the first scientific journal in English in Europe, the *Philosophical Transactions of the Royal Society.* Both were first published in 1665. This is mostly because publishing an article in a prestigious journal implies that the scientific knowledge produced by the authors comply with procedures for assuring high quality not only in science itself but also in scientific discourse. Historically, RAs evolved from the original informative letter written from one scientist to another, and continually developed further to the current form of journal articles.

The overall organization of a RA currently comprises the introduction, methods, results, discussions, and conclusion parts. The abstract was initially not present in a RA, but gradually became an integrated part of it since the 1960's. Nowadays, the abstract is a mandatory component of a RA for most scientific journals. The role of abstracts in RAs has changed: it has become progressively more important within the last few decades. This is mainly due to the explosion in the number of RAs published annually, and also their increasing online availability. Scientists have to select what is worth reading. Such decision is very much influenced by their first contact with the text, i.e., through the authorship, title, and abstract of a RA. Hence, authors convey not only their scientific knowledge in a summarized form in abstracts, but they also want to place themselves and their work reliably in the scientific community through abstracts. Although abstracts are traditionally considered only as a summary or surrogate for a document, they have actually become business cards of their authors.

RAs have been subject of several linguistic studies, mainly in the area of genre analysis (e.g., Banks 2008; Halliday & Martin 1993; Hyland 2004, 2009; Montgomery 1996; Swales 1981, 1990). Contrastively, abstracts "continue to remain a neglected field among discourse analysts" (Swales 1990:

181), sometimes because "space constraints have prevented any investigation of further part-genres such as abstracts [...]" (Swales 2004: 239), although abstracts "are worthy of study because they are significant carriers of a discipline's epistemological and social assumptions" (Hyland 2004: 63).

It is acknowledged that abstracts and RAs differ in their function, linguistic realizations, and rhetorical structure (Lorés 2004: 281). The role of abstracts in scientific knowledge expression goes from "distillation" (Swales 1990: 179), to "act as a report in miniature" (Jordan 1991: 507), and "summary" (Graetz 1982; Kaplan et al. 1994; Ventola 1997) up to "selective representation [...] [of the] exact knowledge of an article's content" (Hyland 2004: 64). Research studies concerning the rhetorical structure of abstracts (e.g., Bondi 2004; Hyland 2004; Liddy 1991; Martín-Martín 2003, 2005; Salager-Meyer 1990), thematic organization (Busch-Lauer 1995; Lorés 2004) and grammatical characterization of abstracts based on the selection of some linguistic features and their analysis over a few selected abstracts (Graetz 1982; Jordan 1991) also contribute to their linguistic characterization. However, with the exception of Bazerman (1984b), who performs a case-study of abstract-RA relationships, all the studies mentioned above focus on abstracts only. They do not compare abstracts to their respective RAs. This gap is one of the major motivations for the present study.

Another prime motivation for this study is how abstracts are to be positioned in a broader linguistic context. While Swales (1990); Hyland (2004) and Swales & Feak (2009) consider abstracts as a "part-genre" of RAs, Jordan regards abstracts as a "special narrow genre within the wider genre of description" (1991: 508). Contrastively, Lorés (2004: 281) describes abstracts as "a genre in its own right which, while sharing many features of the RA, also differs in several important aspects, one of which is its rhetorical structure". However, Lorés focuses only on the rhetorical structure of abstracts not approaching further aspects and differences between abstracts and RAs. Besides, there is not even an "agreement on the concept of genre itself" (Hyland 2009: 26). While Martin approaches genre as "context of culture" (1992a: 495) and as "how things get done, when language is used to accomplish them" (1985: 250), Swales states that a genre "comprises a class of communicative events, the members of which share some set of communicative purposes" (1990: 58) and Hyland understands genres as "schema" (Hyland 2009: 26). These two motivations thus shape the main goals of this research, as discussed in Section 1.1.

1.1 Objectives of the study

This thesis aims to gain insight into the linguistic characteristics of abstracts in direct comparison with their respective RAs and to find differences and similarities between them. Abstracts themselves have been a "rather neglected social artifact of disciplinary life" (Hyland 2004: 83) and a direct analysis comparing abstracts to their RAs has been largely disregarded by present linguistic research (Swales 1990: 181). For this reason, this thesis aims to systematically explore observable linguistic features at both lexical and grammatical levels, and evaluate them qualitatively and quantitatively. This thesis *does not* aim to approach the topic of global organization of abstracts and RAs, e.g., rhetorical structure, thematic organization, inasmuch as these topics have been already expressively covered in the literature (e.g., Bondi 2004; Busch-Lauer 1995; Kaplan et al. 1994; Liddy 1991; Lorés 2004; Martín-Martín 2003, 2005; Nwogu 1993; Ozturk 2007; Saki 2004; Salager-Meyer 1990; Ventola 1997). The investigation of linguistic variation between abstracts and their RAs across disciplines is another pivotal goal of this thesis since different communities may deploy linguistic features in discourse differently (Halliday & Martin 1993; Wignell et al. 1993; Wignell 1998). Finally, based on statistical evaluation of obtained data, this thesis aims to position abstracts and RAs in a broader linguistic context to address the issue on the linguistic relationship between abstracts and RAs.

In order to investigate authentic usage of language, this study is performed over a corpus of abstracts and their respective RAs from scientific journal of several disciplines. The disciplines under study are computer science, linguistics, biology, and mechanical engineering. Mechanical engineering is chosen as a representative for engineering disciplines; biology for natural sciences; linguistics for humanities. Finally, computer science is chosen as a distinctive discipline not fitting perfectly into the classes of scientific disciplines just mentioned. The design, processing, annotation and query of the corpus under study follows the current standards recommended by corpus linguistics methods (e.g., Biber 1990, 1993a; Biber et al. 1998; McEnery & Wilson 2001; Sinclair 1991).

The criteria for the selection of linguistic features for the systematic quantitative analysis of the corpus follows not only preeminent work on corpus-based quantitative linguistic analysis (e.g., Biber 1988, 1993b, 1995, 2006a,d; Biber & Finegan 1994) but is also based on primary data directly obtained from the corpus under study. Lastly, the evaluation of the results is substantiated by current and traditional statistical methods and practices (e.g., Baayen 2008; Baroni & Evert 2008; Gries 2006, 2007, 2008a,b, 2009a;

Manning & Schütze 1999; Oakes 1998).

Furthermore, this work has theoretical underpinnings. As Oesterreicher (2001: 1564) points out, theoretical assumptions are always present in any linguistic analysis. What is needed for this study is a linguistic theory that considers the functional variation of language and the context of situation in which this variation takes place, thereby delivering a systematic analytical framework for lexical and grammatical qualitative and quantitative analysis of linguistics features of this variation. Systemic Functional Linguistics (SFL; Halliday 1985a; Halliday & Hasan 1989; Halliday 2004a) fulfills these needs since the interest in functional variation of language is inherent in SFL (Halliday 2004a: 33ff). Hence, SFL and corpus linguistics (CL; Fillmore 1992; McEnery & Wilson 2001; Sinclair 1991) are the theoretical and methodological underpinnings of this research.

The characteristics of this study can be summarized as follows:

- Objects of study
  - Abstracts and their RAs
  - Source: scientific journals
  - Disciplines: computer science, linguistics, biology, and mechanical engineering
- Theoretical underpinnings
  - Systemic Functional Linguistics
- Methods
  - Corpus linguistics
  - Quantitative analysis of linguistic features at both lexical and grammatical level
  - Statistical evaluation of data
- Issues addressed
  - Differences and similarities on the quantitative distribution of selected linguistic features at both lexical and grammatical level between abstracts and their RAs
  - Relationship between abstracts and RAs in a broader linguistic context

1.2 Organization of the thesis

This first chapter has introduced the topic of this thesis, the linguistic investigation of differences between abstracts and their research articles based on the quantitative analysis of the distribution of selected features. It has also discussed the issues that motivated this study and presented the objectives and organization of this thesis.

Chapter 2 gives an overview of the state of the art on the linguistic research of abstracts and research articles. The first section in this chapter, Section 2.1, presents a survey on the development of RAs and abstracts, followed by a review on the linguistic research on them so far in Section 2.2. Then, Section 2.3 introduces the principles of genre analysis, the field of linguistics within which RAs have frequently been object of study. Section 2.4 summarizes the work performed in the area of register analysis. The next section, Section 2.4.2, presents an overview on SFL, the theoretical framework of this study. Finally, Section 2.5 discusses the terminological and conceptual differences between genre and register, establishing the working assumptions adopted by this study.

Chapter 3 discusses several methods of empirical qualitative and quantitative linguistic analysis, focusing on corpus-based research, i.e., corpus linguistics, in Section 3.2. After an excursus on the connections between corpus linguistics and SFL in Section 3.3, this chapter ends with a description of methods for statistical evaluation of data in Section 3.4.

Chapter 4 introduces the research design of this study. It presents the corpus in Section 4.1, followed by its processing and annotation in Section 4.2. Then, Section 4.3 formulates the hypotheses that will be tested in the empirical analysis. Finally, the linguistic features chosen for the empirical analysis are presented in Section 4.4 including the criteria applied for feature choice.

Chapter 5 presents the results of the empirical analysis and explores the obtained data thoroughly. The following chapter, Chapter 6, discusses the position of abstracts and RAs in a broader linguistic context based on the theoretical underpinnings of this work in relation to the results of the empirical linguistic analysis. It also concludes this work with a summary of the methodology and findings, and outlines some applications of this work as well as some areas of future research.

Chapter 2

State-of-the-art

This chapter initially presents the state-of-the-art linguistic description of abstracts and RAs, provides an overview on their historical development and discusses the most relevant research performed so far in Sections 2.1 and 2.2, respectively. Section 2.3 provides a review of genre analysis (e.g., Swales 1990, 2004), the area of linguistics where most of the studies concerning abstracts and RAs are to be placed. Register analysis, an approach for investigation of linguistic variation, mainly represented by Douglas Biber's work (e.g., Biber 1988, 1995), is described in Section 2.4. Systemic Functional Linguistics (Halliday 1985a, 2004a), the theoretical underpinnings of this study, is then introduced in Section 2.4.2. Finally, Section 2.5 discusses the controversies involving the concepts of genre and register and their applications in linguistics establishing the analytical framework for this study.

2.1 Brief survey on the historical development of research articles

Until almost the middle of the seventeenth century, Latin was the primary language used for all kinds of scientific writing. In January 1665, the first scientific journal in English was published: the *Philosophical Transactions of the Royal Society*. This journal is one of the "most influential record of scientific research during the seventeenth and eighteenth centuries" (Biber & Conrad 2009: 157) because in contrast to earlier conventions, it established the practice of reporting immediate empirical results of the study of nature. The variety of texts published at that time by the *Philosophical Transactions of the Royal Society* goes from the direct exchange of letters between scientists, which was a very customary way of disseminating scientific knowledge at that time (cf. Banks 2008; Montgomery 1996), up to the "experimental essay". This rather recently developed text type was named by the natural philosopher, chemist and physicist Robert Boyle. Boyle's "experimental essay" was supposed to be very structured. It began with a prologue, in which the reasons for performing a certain experiment were presented, followed by a report of the procedures used in the empirical procedures step-by-step, and finalized by discussion upon the results, which in some cases, led to the formulation of a number of hypothesis (Montgomery 1996: 92-95). Boyle's new style of disseminating science gained acceptance by the scientific community at that time, and became very popular. One of his followers was Isaac Newton, who adopted Boyle's emphasis on providing evidence within empirical science. "For registering the birth of scientific English we shall take Newton's *Treatise on Opticks* (published 1704; written 1675-1687). Newton creates a discourse of experimentation [...]" (Halliday 1993a: 57). Halliday does not claim that this was the first scientific document to be written in English at all. He means, however, that Newton's work revolutionized the practices of the scientific discourse, how scientists reported on science. Typical new linguistic constructions in Newton's *Treatise on Opticks* are "*a* causes *x* to happen" or "*b* causes me to think *y*". The original linguistic motifs introduced by Newton in this text are still characteristics of contemporary scientific discourse and involve the description of experiments, by which clause complexes become intricate, grammatical metaphor is beginning to be used, impersonality in scientific writing is brought by the use of passive voice, and abstract nouns are used as technical terms of physics (Halliday 1993a: 57-62). Newton's linguistic innovations began to percolate through the whole scientific community and raised evolutionary processes that lead to modern scientific discourse.

In the last 350 years, the structure of RAs has changed dramatically. For instance, the usual article length fell from ca. 7,000 to ca. 5,000 words from 1890 to 1900. The shortest RAs were found in 1940, with only ca. 5,000 words. After this, the number of words in a typical RA increased again and reached its current average length of ca. 10,000 words in 1980. Concerning the organization of RAs, only ca. 50% of them were formally divided into sections before 1950, which became a regular feature afterwards (Swales 1990: 114). The dense use of references and the intensive practice of quoting previous works came into use only during the twentieth century. The actual established structure of RAs, divided into the sections of introduction, methods or methodology, results, discussions, conclusions, and references is the result of the evolution of scientific discourse over centuries, gradually focusing on research studies, their findings corroborated by evidence, and theoretical relevance to previous work, written in a strict format (Biber & Conrad 2009; Montgomery 1996; Swales 1990). Abstracts were only introduced into this format during the 1960s, primarily in RAs from medical disciplines (Swales & Feak 2009: 1). Although RAs changed immensely through history becoming more narrowly defined in terms of textual and structural conventions, it preserved its original goal of conveying the results of scientific investigation (Biber & Conrad 2009: 166).

The RA is the product of a long process. A process that has started with doing science, going through several steps of manuscript writing, manuscript submission to a journal, peer-reviewing, revising and sometimes re-writing of the manuscript, until finally putting the process to an end by having the RA published in the journal of choice. Publishing articles in journals became prestigious, especially because of the high quality standards of scientific practices established over time.

Since an abstract is only a distillation of the whole text, it has initially played a secondary role in RAs. The word *abstract* means "a summary or epitome of a statement or document"[2] and has been used in written texts since 1528 (OED Online 1989). Nowadays, however, abstracts function as independent discourses (van Dijk 1980). They have to persuade readers to read the whole RA ahead of them and to convince them that the authors have credibility to address such topic within the scientific community (Hyland 2004: 63-65). At the present time, scientists are flooded with the enormous number of RAs being published daily, and also electronically. Thus, it is of vital importance to select what is worth reading. This selection is done upon evaluation of the journal title, the authorship of the article, its title and abstract, not necessarily in this order. The importance of abstracts in

[2] "abstract, B.n.2" The Oxford English Dictionary. 2nd ed. 1989. OED Online. Oxford University Press. 4 Apr. 2000 <http://dictionary.oed.com/cgi/entry/50000886>.
2. spec. A summary or epitome of a statement or document. Also attrib.
1528 GARDINER in Pocock Rec. Ref. I. I. 117 We send herein enclosed, abstracts of such letters as hath been sent to the pope's holiness. 1715 BURNET Hist. own Time (1766) II. 82 I will give you here a short abstract of all that was said. 1799 WELLINGTON Lett. (G.D.) I. 34 In the abstracts, it appears that the strength of the..forces consisted of 48,000 men. 1863 COX Inst. of Eng. Govt. Pref. 8 Copies or abstracts of State papers and records. 1867 SMYTH Sailors' Word-Bk. s.v. An abstract log contains the most important subjects of a ship's log. 1927 [see ABSTRACTOR]. 1959 L. M. HARROD Librar. Gloss. (ed. 2) 12 Abstract. I. A form of current bibliography in which contributions to periodicals are summarized... When published in periodical form they are known as journals of abstracts. 2. The individual entry. 1962 Lancet 19 May 1068/1 Have you ever tried doing abstracts? I once did – for about a year. It was the American articles that caused me the most anguish.

RAs increased considerably within the last few decades. While RAs have been object of numerous linguistic studies, abstracts have remained a rather neglected field among linguists. Previous linguistic research involving RAs and abstracts are discussed in Section 2.2.

This brief survey on the historical development of RAs is necessarily incomplete. Comprehensive information on this topic can be found in the works of e.g., Banks (2005a, 2008); Bazerman (1988); Biber & Conrad (2009); Halliday & Martin (1993); Halliday (2004b); Hyland (2009); Martin & Veel (1998); Meadows (1980); Montgomery (1996); Randaccio (2004); Swales (1990, 2004).

2.2 Linguistic analysis of research articles and abstracts

This section presents an overview on previous linguistic research on RAs and abstracts, in Sections 2.2.1 and 2.2.2, respectively. This overview, however, does not aim to provide a full inventory of previous linguistic studies on abstracts and RAs; but it rather aims to cite and discuss the main previous works in different areas of linguistics.

2.2.1 Research articles

An extensive number of linguistic studies has been carried out on RAs as a whole. Some of the studies are concerned with the diachronic analysis of RAs. For instance, Bazerman (1984b, 1988) traces the development of experimental articles in English in the *Philosophical Transactions of the Royal Society*. His pioneering studies provide a comprehensive bibliography and index, making them an outstanding introduction to the work being done in history of science. In another work, Bazerman investigates the textual development of RAs in the *Physical Review* over the last century. He concludes that "this period marks the rise of American physics from backwardness to world dominance, reflected by the journal's rise from a local university organ to the primary international journal of physics" (1984a: 166). The results of his study show how RA's properties, e.g., article length, references, syntactic and lexical features, and organization changed over time. Another example of diachronic study on RAs is the work of Atkinson (1992), who discusses the evolution of medical writing based on changing language and rhetoric of medical research reporting published in the oldest continuing medical journal in English, the *Edinburgh Medical Journal*. Moreover, Salager-Meyer

(1999) examines the diachronic evolution of referential behavior in medical written-English discourse in a corpus of 162 medical articles published in 34 British and American medical journals between 1810 and 1995. The use and frequency of reference patterns over the years indicates the shift from a non-professionalized, privately and individually-based medicine to a professionalized and specialized medicine, a technology-oriented medical research and a highly structured scientific community. Medical sciences, however, is not the only discipline to have their RAs studied. For instance, the historical development of RAs on the physical sciences and their lexico-grammatical innovations is the focus of Halliday (1993a). The research focus of Wignell's work lies on the linguistic analysis of RAs in social sciences and geography (Wignell et al. 1993; Wignell 1998, 2007), which shows the domain specific differences in the linguistic realization of scientific discourse. Biber & Finegan (1989) report on a comparative multi-dimensional analysis of the linguistic development of functionally different registers, comparing essays, fiction and personal letters. Biber & Finegan (1992) present a comparative diachronic analysis of written and speech-based genres including scientific writing. Banks (1991, 1994, 2005a,b, 2006, 2008) is not only interested in the historical development of RAs, but also in specific linguistic features, e.g., nominalization, passive voice, personal pronouns, and lexical hedging. His work is mainly exploratory; hence, computer-aided analysis are often deliberately excluded. Finally, corpus representativeness for diachronic linguistic studies based on a corpus containing texts from the *Philosophical Transactions of the Royal Society* from the seventeenth century is recently discussed by Moessner (2009).

Lexico-grammatical aspects of RAs have also been the focus of numerous linguistic research so far. Gerbert (1970) analyzes the use of verbs in English technical writing. He concludes that the present tense is usually used in representing definitions, descriptions, and observations, while perfect tense is mostly used in describing research processes. Inman (1978) and Love (1993) investigate the distribution of lexical items in RAs from different disciplines. Salager-Meyer (1994) discusses how the communicative purposes of the different rhetorical sections of medical RAs influence the frequency of hedges used in each section. Her results show that the choice of hedging is imposed by the general structure and communicative purpose of the discourse. Conrad (1996) investigates numerous lexico-grammatical features in several academic texts from biology, showing, for instance, that RAs have a more informational focus and impersonal style than textbooks. Biber et al. (1998) report on the analysis of grammatical features of RAs of several disciplines, comparing ecology articles with history ones, among others. They show, for example, that the first ones are characterized by more impersonal features (agentless passives, conjuncts, etc.) and the latter by more narrative features (past tense verbs, present particle clauses, etc.). Hyland (1998) demonstrates that RAs make wide use of hedges sometimes even more frequently than modals to express uncertainty. The preference of RAs for interpersonal discourse through hedges, personal and frame markers is also corroborated by Hyland (1999). Furthermore, Hyland (2002) shows that the use of directives to guide readers through the text is very common in RAs of hard sciences and relatively unusual in social sciences. Gledhill (2000a,b) studies the structure of collocations as lexico-grammatical patterns and their discourse functions in RAs comprehensively. Moreover, Marco (2000) reports on collocational frameworks in medical RAs. In a recent work, Hyland reports on the forms, structures and functions of word clusters in a corpus of research articles, doctoral dissertations, and master's theses. He shows that the study of clusters is an appropriate indicator to gain insights into "the ways writers employ the resources of English in different contexts, and with the potential to inform advanced academic literacy instruction" (Hyland 2008: 60).

A further focus of the linguistic research on RAs is their rhetorical organization and thematic structure. Swales (1981) studies the introductory parts of RAs in detail. He establishes rhetorical moves in the argument structure of such introductions, and aims to contribute pedagogically to native and foreign language communication skills teaching. His research remains one of the most detailed rhetorical analysis of RAs so far, especially in the books from 1990 and 2004, where he claims that there is a fundamental rhetorical system and a stereotypical rhetorical structure in RAs. He introduces the terms *moves*, which are obligatory, and *steps*, which are optional, as part of the rhetorical structure of RAs so that the desired argumentation can be enfolded through the text. Swales's model of textual macrostructure of RAs has been greatly accepted within the linguistic community and has been widely further adopted. Nwogu (1991, 1993, 1997) intensively studies the function and structure, theme-rheme patterns, paragraph development and rhetorical moves in RAs of reputable medical journals. Hunston (1993)'s study provides insights into the relationship between the evaluative expression of RAs and ideology of science in general. More recently, Ozturk (2007) studies the textual organization of RA introductions in the discipline of applied linguistics, which explores sub-disciplinary variation in the move structures of the texts under study.

The studies presented in this section do not represent the whole spectrum of linguistic analysis on RAs. Additional information on this topic can be found in e.g., Biber & Conrad (2009); Gledhill (2000b); Hyland (2009); Swales (2004).

2.2.2 Abstracts

Linguistic research on abstracts has been mainly descriptive as opposed to linguistic research on research articles. One of the first linguistic analysis of abstracts was performed by Graetz (1982), who reported on a study over 87 abstracts from the disciplines of health sciences, social sciences, education, and humanities. She aims to gain insights into their linguistic properties. Her aim was to improve teaching practices to students of English as a foreign language, so that they can successfully extract structural information from abstracts. She defined that the purpose of abstracts was to "give the reader an exact and concise knowledge of the total content of the very much more lengthy original, a factual summary which is both an elaboration of the title and a condensation of the report" (Graetz 1982: 23). Furthermore, she argued that the language of abstracts is characterized as follows:

> The abstract is characterized by the use of past tense, third person, passive, and non-use of negatives. It avoids subordinate clauses, uses phrases instead of clauses, words instead of phrases. It avoids abbreviation, jargon, symbols and other language shortcuts which might lead to confusion. It is written in tightly worded sentences, which avoid repetition, meaningless expressions, superlatives, adjectives, illustrations, preliminaries, descriptive details, examples, footnotes. In short it eliminates the redundancy which the skilled reader counts on finding in written language and which usually facilitates comprehension. (Graetz 1982: 23)

Her work, although pioneer, has been often criticized as being "a little bold" (Swales 1990: 180) and for the fact that "it is easy enough to find counter-examples" Hyland (2004: 65). Moreover, according to Ventola (1997: 345), Graetz's classification criteria is "relatively ad hoc [and] it is merely a list of some of the realizations found in the scientific abstracts studied".

However, there are many other relevant studies on lexico-grammatical features of abstracts. One example is Fluck (1988), who quantitatively analyzes linguistic features of abstracts of economics, linguistics, and metal industry in German. The results of this study indicate that abstracts are characterized by complex nominalizations, extensive noun compounding, impersonality, use of third person, passive voice and present tense. Furthermore, Gnutzmann (1991) compares passive voice use quantitatively in abstracts and in conclusions of RAs in English and German from the disciplines of linguistics, sociology, and theoretical engineering. His results showed that abstracts use passive voice more frequently than conclusions.

There are very few linguistic analysis of abstracts that are compared directly to their RAs (cf. Swales 1990: 181). One of these is Bazerman (1984b), who investigates a case-study on the construction process of RAs. However, this work is just exemplary since it is based on archival manuscripts from the physicist Cromptom in 1925. The other known study comparing abstracts to their whole texts is the work done by Kretzenbacher (1990), who examines a corpus of 20 RAs from the humanities in German. Kretzenbacher reinforces the general finding that abstracts have a more nominal style, e.g., higher noun-per-sentence ratio and nominalizations, use genitive attributes and definite articles more often than the correspondent RA. In contrast, RAs tend to use modal verbs more frequently than their abstracts. Jordan (1991) defines two types of abstracts, the descriptive and the informative ones, aiming to provide linguistic criteria for the distinction between them. However, this analysis is performed over a very small number of abstracts and the criteria for abstract classification concentrate on the use of passive voice and verb tenses. Ventola (1994) reports on textual and syntactic analysis concerning the problems of writing of abstracts in a foreign language, in this case, English. Later on, Ventola (1997: 349) argues that "abstracts should be taken as a serious object of linguistic study" and provides a comprehensive overview of the linguistic analysis of abstracts up to then. Ventola claims that abstracts should not only be the object of theoretical studies but also how important the application of such studies as well as the cooperation between applied and field experts is for scientific writers. Dorgeloh & Wanner (2003) studies the representation of agentivity in abstracts of RAs through the classification of the verbs used in several categories, e.g., reporting, mental, relational verbs. Their results indicate that there is no generalization to be made concerning a possible loss of the agentivity in scientific discourse. Hyland (2004) analyzes how authors claim credibility and promote themselves in abstracts of a multidisciplinary corpus. The major criticism to such studies so far is that they are mainly descriptive, free of theoretical rooting and very often exemplarily.

Lorés (2004) analyzes abstracts from RAs according to their rhetorical organization and thematic structure from the discipline of linguistics. Her research follows the classification and analysis methods of Swales (1990) and focuses especially on the thematic progression of the abstracts under study. Thematic progression patterns in abstracts is also the focus of Saki

(2004)'s work, while their discourse structure is investigated in Liddy (1991). Salager-Meyer (1990, 1992) performs a corpus-based study of verb tense and modality distribution in medical abstracts and examines how the meaning conveyed by the different tenses and verbs is related to the function of the different rhetorical divisions of abstracts. Her findings show that the active past tense is the most frequent verb form followed by the past passive. The results also indicate a correlation between tense and form of verbs distribution with subsections and rhetorical moves within abstracts structure. For instance, *may* is most frequent in the conclusion part of abstracts, whereas *should* is mostly used in *recommendation* moves in abstracts, and past tense is mainly used in moves of *statement of problem* and *data synthesis*. Salager-Meyer's research is consistent with the rhetorical model of text analysis and moves categorization developed by Swales (1990). Moreover, Stotesburry (2003) performs an evaluation of the use of stance expressions in abstracts in the humanities, social and natural sciences, showing that abstracts in the humanities tend to use more evaluative expressions, while abstracts in the natural sciences prefer modal verbs. Lastly, expressions of epistemic modality are more common at the end of abstracts from natural sciences, while they are more frequent at the results part of abstracts of social sciences.

Another aspect of the linguistic research on abstracts is herewith addressed. These are studies that compare traditional abstracts with a recently new form of abstracts, the structured ones. Structured abstracts contain sub-headings, such as background, aim, method, results, and conclusions. Such abstracts are mostly found in medical journals. However, they have been increasingly gaining acceptance in recent years, also in other disciplines, such as economics. Hartley et al. (1996) investigate whether structured abstracts may have an additional advantage of being easier to search in comparison to traditional ones. Their findings support the initial hypothesis that it is easier for the readership to search information in structured abstracts. In another work, Hartley & Sydes (1997) compare the readability of structured abstracts with traditional ones. The results of this work indicate that structured abstracts are not always easier to read than the traditional ones. Finally, Hartley (1999) investigates whether structured abstracts might be appropriate for the journal *Applied Economics*. His work is exemplary considering only a few abstracts. The measures performed are word length, information content, readability through a computer-based readability score, and reader preferences which were collected by asking research fellows. His conclusions indicate that the structured abstracts are usually longer, more informative and found to be clearer by the their readership. This supports Hartley's previous view that structured abstracts are more effective than traditional ones.

One last aspect of the linguistic research on abstracts involves the question on how abstracts and RAs relate in a broader linguistic context. Swales (1990); Hyland (2004) and Swales & Feak (2009) consider abstracts as a "part-genre" of RAs[3]. Similarly, Jordan views abstracts as a "special narrow genre within the wider genre of description" (1991: 508). Conversely, Lorés (2004: 281) describes abstracts as being "a genre in its own right which, while sharing many features of the RA, also differs in several important aspects, one of which is its rhetorical structure". However, her study focuses only on the rhetorical structure of abstracts not approaching further aspects and differences between abstracts and RAs. Apart from that, there is even no "agreement on the concept of genre itself" (Hyland 2009: 26). While Martin approaches genre as "context of culture" (1992a: 495) and as "how things get done, when language is used to accomplish them" (1985: 250), Swales affirms that a genre "comprises a class of communicative events, the members of which share some set of communicative purposes" (1990: 58) and Hyland understands genres as "schema" (Hyland 2009: 26).

This brief overview on previous linguistic research on abstracts and research articles showed that the majority of them is descriptive and that they do not compare abstracts directly to their RAs. Moreover, current controversies concerning the relationship between abstracts and RAs in the linguistic context were presented and discussed. The aims of this study are, therefore, twofold. First, it aims to fill the gap concerning the lexico-grammatical analysis of abstracts in direct comparison to their RAs in English. Second, it addresses conceptual and terminological uncertainties, in order to clearly position abstracts and RAs in a broader linguistic context. For this reason, this study is rooted on Systemic Functional Linguistics. SFL is a social semiotic approach to language centered around the notion of language function, which is suitable as a theoretical background for this research. However, before SFL is introduced and discussed in details in Section 2.4.2, two other relevant linguistic approaches to research on scientific discourse, genre analysis and register analysis, are presented in Sections 2.3 and 2.4, respectively.

[3] "*Genre* is a name for a type of text or discourse designed to achieve a set of communicative purposes. Following this terminology, the research article is a genre, and various parts of it, such as the Abstract and Discussion, are part-genres" (Swales & Feak 2009: 1).

2.3 Genre analysis

Genre analysis can be defined as an umbrella term, covering "a range of tools and attitudes to texts, from detailed qualitative analysis of a single text to more quantitative counts of language features" (Hyland 2009: 25). It is focused on the notion of *genre*. Notwithstanding the fact that there is still no agreement on the definition of the concept of genre (cf. Bawarshi & Reiff (2010: 3), Biber & Conrad (2009: 21-23), Hyland (2009: 26)), a generally acknowledged definition of *genre analysis* that can be considered is the one from Bhatia, who defines genre analysis as "the study of situated linguistic behaviour in institutionalized academic or professional settings" (Bhatia 2002: 22).

The term *genre* is multifarious; it has different meanings in different research areas, e.g., folklore studies, literary studies, linguistics, rhetoric. One possible definition of *genre* according to The Oxford English Dictionary[4] is "a particular style or category of works of art; esp. a type of literary work characterized by a particular form, style, or purpose". However, according to such definitions, genre keeps being "a fuzzy concept, a somewhat loose term of art" (Swales 1990: 33). In the last few decades, major linguistic work on genre has been performed by Swales (1981, 1990, 2004) among others, e.g., Bazerman (1994); Hymes (1974); Saville-Troike (1982). In one of his earliest works, Swales defines genre as "a more or less standardized communicative event with a goal or set of goals mutually understood by the participants in that event and occurring within a functional rather than a social or personal setting" (Swales 1981: 10). In later work, he provides a more detailed working definition of genre, although still recognizing that "there remain several loose ends" (Swales 1990: 57):

> 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 the choice of content and style.
>
> (Swales 1990: 58)

In one of his most recent works, Swales (2004: 61) admits that his former definition of genre is not applicable in the analysis of all cases and even

[4] "genre" The Oxford English Dictionary. 2nd ed. 1989. OED Online. Oxford University Press. 4 Apr. 2000 <http://dictionary.oed.com/cgi/entry/50093715>.

often prevents the recognition of newly explored or emerged genres. For this reason, he now suggests that genre is to be regarded metaphorically, as *frames* or *schema* for social action guiding users to achieve particular purposes through language. According to this definition, academic discourse would be a "constellation" of written (e.g., RAs, conference abstracts, PhD Dissertations, grant proposals, textbooks, book reviews, etc.) and spoken (e.g., lectures, seminars, colloquia, office hour meetings, PhD defenses, etc.) genres (Hyland 2009; Swales 2004).

Genre analyses of academic discourse, including RAs and abstracts, based on Swales's work and genre definition, have been widely performed, especially in the areas of English for Special Purposes (ESP) and rhetorical analysis (cf. Section 2.2.1 and 2.2.2). A typical genre analysis starts with the identification of the genre within a discourse community, followed by the description of its intended communicative purpose. Then, the analysis examines the genre's organization, i.e., its schematic structure according to the categories of the rhetorical *moves* defined by Swales. Additionally, textual and linguistic features realizing such rhetorical moves are also often analyzed. Nonetheless, the main focus of the genre analysis approach described here lies on making pedagogical improvements on ESP teaching for both native and non-native speakers of English (Swales 1990: 232). The major criticism of genre analysis approach, besides the fact that there is still no general agreement on the definition of *genre* itself, is its extreme focus on singular rhetorical events and strategies and its lack of underpinnings for a general theory of language (e.g., Bhatia 1993). Moreover, the emphasis on the direct transmission of text types implied by such an approach "does not necessarily lead on to a critical appraisal of that disciplinary corpus, its field or its related institutions, but rather may lend itself to an uncritical reproduction of discipline" (Luke 1996: 314). Thus, Hyland (2003: 25) argues that "teaching genres may only reproduce the dominant discourses of the powerful and the social relations which they construct and maintain".

There are other approaches or perspectives in linguistically analyzing texts, for instance, register analysis and Systemic Functional Linguistics, which are more suitable for the current study. Register analysis is based on the concept of *register*, the functional variation of language, and is concerned with the qualitative and quantitative analysis of lexico-grammatical features of texts and text excerpts. Register analysis is discussed in detail in Section 2.4 and representative research in this area is given by Biber (1988, 1995); Biber & Finegan (1994); Ghadessy (1993); Martin (1992a, 1993); Ure (1971, 1982), among others. Systemic Functional Linguistics is developed by M. A. K. Halliday (Halliday 1985a, 2004a; Halliday & Hasan

1989), among others, and provides a comprehensive theory of how language functions to make meaning in a socio-cultural context. This theory is presented and discussed in detail in Section 2.4.2. It constitutes the theoretical underpinnings for this work.

2.4 Register analysis

The notion of *register* is a long-established concept in linguistics. According to The Oxford English Dictionary, the term *register* was first introduced in linguistics by Reid (1956: 32).

> He will on different occasions speak (or write) differently according to what may roughly be described as different social situations: he will use a number of distinct 'registers'. OED Online (1989)[5]

By this time, studies of language focused on its structure, e.g., by identifying structural units and classes of given language and describing how these units would combine to create larger structures, and not on the variation of language. Generally, variation of language can be divided into two types, *user*-dependent and *use*-dependent variation (Gregory 1967). Examples of *user*-dependent variation are dialects, sociolects, and genderlects. Contrastively, the *use*-dependent variation of language is exemplified in the language of science and technology, legal English, language of weather reports and recipes, among others. When compared to *dialects*, which can be seen as *what* a person speaks, determined by *who* he is, *register* is *what* the person is speaking, determined by *what* he is doing (Webster 2009: 445). Thus, the term *register* has been used as an equivalent to language variation according to use. According to Beaugrande (1993), similar concepts to register are already proposed by Pike (1967), i.e., "the universe of discourse" and Firth (1957, 1968), i.e., "restricted language". Ure (1969a,b) further developed the concept of register as "situationally-differentiated language variety" (1969a: 107). However, it was M. A. K. Halliday who coined and spread the notion of register as used in contemporary linguistics. In an early work, Halliday & Hasan (1976: 22) defined register as "the linguistic features which are typically associated with a configuration of situational features". According to this definition, the more specifically the context of situation can be described, the more specifically the properties of text in

[5] "register, n.1" The Oxford English Dictionary. Draft Revision Mar. 2010. OED Online. Oxford University Press. 4 Apr. 2000 <http://dictionary.oed.com/cgi/entry/50201234>.

such a situation can be predicted. Thus, register is a setup of semantic resources that any member of a culture typically associates with a given situation.

> Types of linguistic situation differ from one another, broadly speaking, in three aspects: first, as regards what actually is taking place; secondly, as regards what part the language is playing; and thirdly, as regards who is taking part. These three variables, taken together, determine the range within which meanings are selected and the forms which are used for their expression. In other words, they determine 'register'. (Halliday 1978: 31)

As stated by Halliday & Hasan (1976), these three aspects mentioned above are called *field*, *mode* and *tenor* of discourse, respectively. More specifically, the field of discourse represents the total event in which language is functioning, which includes not only the topic of the text, but also the purposive activity of the speaker or writer. The mode of discourse can be defined as the function of the text in this event, comprising the channel, e.g., written, spoken, written-to-be-spoken, and rhetorical mode, e.g., narrative, persuasive, didactic. Finally, the tenor of discourse reflects the type of role interaction between the participants involved in the event, i.e., their social relations, e.g., expert-to-expert, expert-to-lay-person, etc. Thus, the main purpose of register analysis is to find out "what situational factors determine what linguistic features" (Halliday 1978: 32).

> A register is a semantic concept. It can be defined as a configuration of meanings that are typically associated with a particular situational configuration [...]. But since it is a configuration of meanings, a register must also, of course, include the expressions, the lexicogrammar and the phonological features, that typically accompany or REALISE these meanings. (Halliday & Hasan 1989: 39)

A given context of situation corresponds, therefore, to a certain register, i.e., a semantic variety in the linguistic system, realizing its own configurations of lexico-grammatical features. Figure 2.1 shows different contexts of situation corresponding to different semantic systems, i.e., registers, which are colored grey. These different registers are realized in the several ways by the lexico-grammatical system of language. This means that the semantic level can be seen as a repertoire of situation-specific semantic systems, including different text structures associated with different situations, and being all realized by the "one highly generalized grammatical system", as

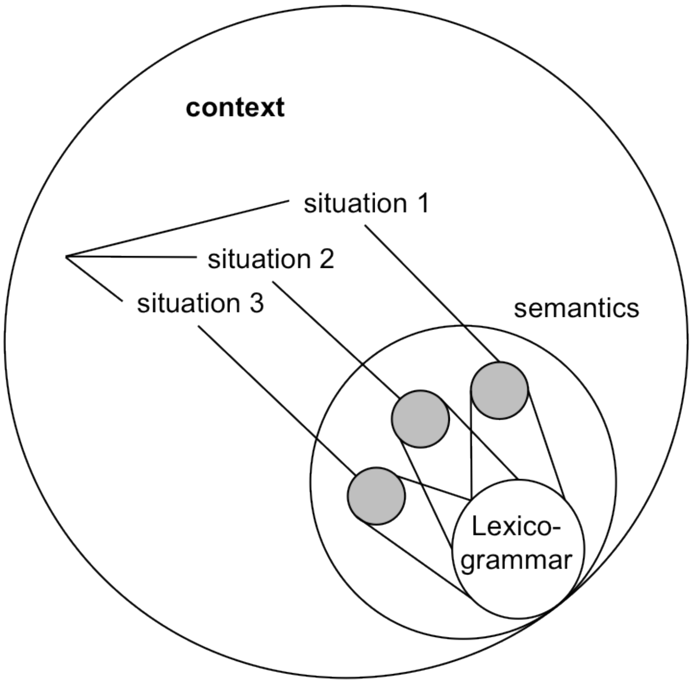


Figure 2.1: Context of situation & register (Matthiessen 1993: 253)

stated by Matthiessen (1993: 253), who provides a comprehensive survey on the development of register theory.

Register analysis has developed quickly in the last decades and several researchers have adopted this approach for analyzing genuine texts and establishing their characteristic linguistics features. This approach can be used both for diachronic as well as for synchronic linguistic analysis of texts. One example of diachronic register analysis is the work of Halliday (1988) in the evolution of the language of physical science. It shows the development of linguistics features characterizing such a register, e.g., nominalization, grammatical metaphor, etc. Most of the studies are, however, synchronic, investigating how written and spoken language is used in various contexts of situations by identifying linguistic features in genuine texts realizing these different registers, (e.g., Ghadessy 1988, 1993, 1999; Halliday & Martin 1993; Martin 1983; Martin & Veel 1998; Neumann 2003, 2008; Steiner 1996; Teich 2003; Ure 1971, 1982; Ventola 1992, 1994, 1996, 1997; Wignell 1998, 2007; Wignell et al. 1993).

2.4.1 Multi-dimensional approach to register variation

Register analysis has profited immensely from the developments in computational linguistics in the last few decades. It has also allowed researchers to gather, process, and analyze a greater number of genuine texts than in the past. One of the most prominent work, diachronic as well as synchronic, in register analysis of large numbers of texts is the research of Douglas Biber and his collaborators (e.g., Biber 1988, 1993b, 1995, 2006a,d; Biber & Conrad 2009; Biber & Finegan 1989, 1992, 1994). He proposes a comprehensive analytical framework for quantitatively analyzing register and register variation. Based on Halliday's definition of register (1978: 31; cf. p. 20), Biber argues that such analytical framework should provide tools for the identification, quantification and classification of the three typical components of register analysis: the situational, and the linguistics characteristics of register, and the functional associations between these two. According to Biber (1994: 35), such analysis are inevitably quantitative since "register distinctions are based on differences in the relative distribution of linguistic features, which in turn reflect differences in their communicative purposes and situations". As Halliday (1988: 162) points out, register can also be defined as "a cluster of associated features having a greater-than-random [...] tendency to occur". Based on the notion of linguistic co-occurrence Biber develops a *multi-dimensional approach to register variation*, by which different patterns of co-occurrence of linguistic features are analyzed as underlying dimensions of functional variation. One of the major distinguishing aspect of Biber's framework is that it considers "register variation as *continuous* rather than discrete" (Biber 1994: 36; emphasis added). Hence, the focus of his multi-dimensional approach is on the relative distribution of common linguistic features, i.e., co-occurrence patterns of *register markers*, flowing across register variation. In the preliminary steps in the development of this approach, he identified 67 linguistic features[6], i.e., register markers (e.g., lexical classes, grammatical categories, syntactic constructions), that may have a functional association in texts, through the quantitative analysis of these features over a large number of naturally occurring texts. These register markers were then organized into 16 major grammatical and functional categories, e.g., *nominal forms*, (e.g., nominalizations, gerunds and total other nouns), *pronouns and pro-verbs*, *passives*, *lexical specificity* (e.g., type/token ratio and mean word length), *modals*, etc. The co-occurrence patterns of these grammatical and func-

[6] A complete list of all register markers used in the multi-dimensional approach can be found in Conrad & Biber (2001: 18-19).

tional categories were then grouped into seven *factors*. This has been done using a statistical technique known as factor analysis. Finally, these *factors* were interpreted as seven *dimensions of variation* used for register comparison. Thus, the first of the seven dimensions, which represent a continuum along which registers may differ, is called *Involved vs Informational Production*, by which high frequencies of occurrence of first- and second-person pronouns, wh-questions, amplifiers are interpreted as an indication of interpersonal interaction, i.e., a higher involved text production. Contrastively, high frequencies of nouns, prepositional phrases, type/token ratio, and attributive adjectives indicate a more informational focus in the text production. The second dimension in this approach is *Narrative vs Non-narrative Discourse*. Linguistic features contributing to the positive characterization of narrative registers, e.g., fiction prose, are past tense verbs, third-person pronouns, synthetic negation, and present participial clauses, among others. Non-narrative registers, such as academic discourse and news, have lower frequency of occurrence of such linguistic features. The third dimension is called *Elaborated vs Situation-dependent Reference*. Linguistic features, contributing to a more elaborated discourse, are, for instance, phrasal coordination, nominalizations, wh-relative clauses, which are highly frequent in, e.g., academic discourse. Time and place adverbials and adverbs in news registers are features with high frequency of occurence that indicate a more situation-dependent register. *Overt Expression of Persuasion / Argumentation* is the name of the fourth dimension. Features contributing to a higher expression of persuasion / argumentation are modals, suasive verbs and infinitives, among others. These occur highly in registers such as professional letters and editorials. Contrastively, news registers are not overtly argumentative, showing lower frequency or even absence of these features. The fifth dimension is *Abstract vs Non-abstract Style*. Similarly to dimensions 2 and 4, it has only positive loadings, e.g., conjuncts, passives, adverbial subordinators, etc. While academic discourse and official documents show a high frequency of these features, conversation and fiction show practically the absence of them. This confirms the expectation for academic discourse being much more abstract than other registers. The last two dimensions, dimension 6, *On-line Informational Elaboration Marking Stance*, and dimension 7, *Academic Hedging*, are the most difficult ones to interpret (Conrad & Biber 2001: 39). These have few features with important loadings, and have been less used in register analysis research. Particullarly the interpretation of dimension 7 still needs to be rectified by further research. Typical features contributing positively for dimension 6 are *that*-complement and -relative clauses, whereas down-toners, adverbs, and attributive adjectives are important for dimension 7.

Biber's multi-dimensional approach for register analysis is, therefore, a comparative perspective, where patterns of register variation are quantitatively investigated. Moreover, it is not rooted in any specific theoretical framework. When a large quantitative unstructured feature set is statistically processed, it will allow English teachers to produce better learning materials.

> In the field of ESP - English for Specific Purposes - researchers and practitioners seek to understand the linguistic characteristics of specialized registers in English. One major goal of such research is to design the best possible materials and activities to help students comprehend and produce these registers appropriately.
>
> (Biber et al. 1998: 157)

This approach requires no hypothesis formulation prior to the experiments and provides a substantial overview over register variation. One of the many advantages of this approach is precisely the fact that no hypothesis is required prior to the quantitative investigation of linguistic features because it allows linguists to gain insights into the variation of many different registers at once. However, this approach has been criticized for relying strongly on statistical techniques, which are themselves not faultless, as well as for its lack of rooting in a broader linguistic theory, e.g., one which considers language entirely as a social-semiotic system. Systemic Functional Linguistics is a linguistic theory that is used as the theoretical underpinnings of this research, and is discussed in the next section, Section 2.4.2.

2.4.2 Systemic Functional Linguistics

The theory of language used as a basis here is the Systemic Functional Linguistics (SFL; Halliday 1985a, 1985b; Halliday & Hasan 1989). SFL treats language use as being inherently context-dependent, giving rise to registers, i.e., patterns of language according to use in context. Hence, *register* is defined in SFL as "what you are speaking at the time, depending on what you are doing and the nature of the activity in which language is functioning." (Halliday & Hasan 1989: 41). SFL is considered with form and function of language as well as the role of context in human communication, thereby providing an analytical framework for lexical and grammatical qualitative and quantitative analysis of linguistic features of language variation.

> A characteristic of the approach we are adopting here, that of systemic theory, is that it is comprehensive: it is concerned with language in its entirety, so that whatever is said about one aspect is to be understood always with reference to the total picture.
>
> (Halliday 2004a: 19)

SFL established a multidimensional model for the description of languages and its architecture evolved continuously since 1970s (e.g., Halliday 1959, 1985a; Halliday & Hasan 1989; Halliday & Matthiessen 2006; Martin 1992a; cf. Matthiessen 2007 for a survey on the historical development of SFL). Currently, SFL's multidimensional model follows the parameters of *stratification*, *metafunction* and *instantiation* (Halliday 2004a; Martin 2007; Teich 2003). The first parameter, *stratification*, means that language is "a complex semiotic system, having various levels, or *strata*" (Halliday 2004a: 24). According to Martin (1992a), SFL distinguishes between the strata of phonology, lexicogrammar, semantics, register, and more recently, genre, and ideology. Initially, there was only the concept of register and there was no need for genre or ideology in the SFL model. Both were added later on by Martin (1992a) (cf. Halliday 1978; Halliday & Hasan 1976, 1989). Figure 2.2 shows the strata of language as in SFL's model through "the metaphor of concentric circles" (Martin 1992a: 496), by which the larger circles recontextualize the smaller ones and their sizes indicate that each strata becomes a larger unit, from phonology to ideology. This means that at the strata of phonology, the focus of analysis is on syllables and phonemes, at the strata of lexicogrammar it lays on the clause, and at the semantic strata the focus of analysis lays on paragraphs. Additionally, the focus at the level of register lays on "a stage in a transaction", at the level of genre on whole texts, and at the level of ideology on "discourses manifested across a range of texts" (Martin 1992a: 496).

The second parameter of SFL's multidimensional model, *metafunction*, relies on the multifunctional nature of language, thereby distinguishing three functions or meanings: the *ideational*, the *interpersonal* and the *textual. Ideational*, the first metafunction that is further classified into logical and experiential, is related to the construction of institutional activity, i.e., to the construction of human experience (Halliday 2004a; Martin 2009). The ideational metafunction is linguistically realized in the *field* of discourse, which refers to what is happening, i.e., to the nature of the action taking place in the discourse. The parameter of field characterizes texts in terms of their domain-specificity, as described in terms of lexis, specialized terminology, etc. The second metafunction, *interpersonal*, is related

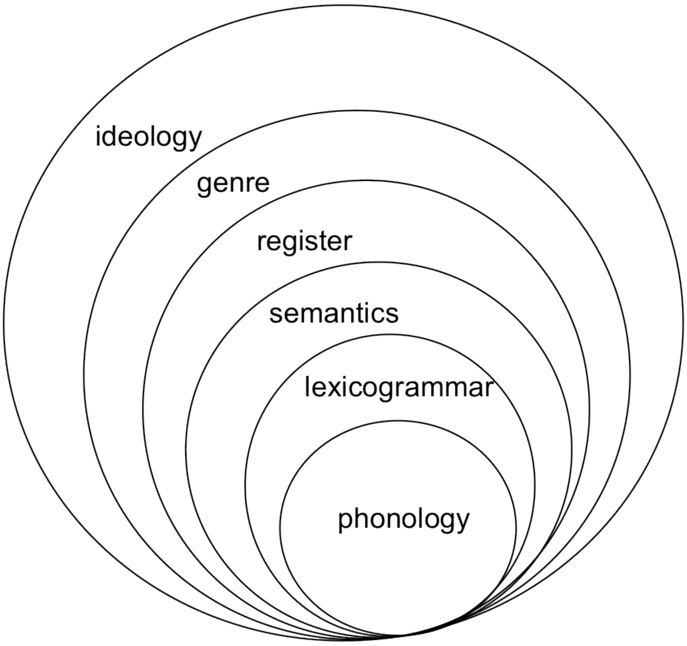


Figure 2.2: Stratification: Language & its semiotic environment (adapted from Martin 1992a: 496)

to the negotiation occurring in the social action. The interpersonal metafunction is linguistically realized in the *tenor* of discourse, characterizing texts in terms of the interaction between the participants involved in the discourse situation, e.g., expert-to-expert for abstracts and RAs (Martin & Rose 2007). Finally, the third metafunction, is called *textual*. It reflects how information flows and creates cohesion and continuity within discourse (Halliday 2004a). Moreover, it is also linguistically realized in the *mode* of discourse, i.e., in the symbolic organization of discourse. The parameter of mode refers to the realization of the communication process in terms of channel and medium. For abstracts and RAs, for example, the channel is indirect, i.e., non-face-to-face communication, and the medium used in the communication is written-to-be-read (Halliday & Hasan 1989). Taken together, the parameters field, tenor, and mode of discourse constitute the *register* of a discourse. In other words, *register* refers to “the semiotic system constituted by the contextual variables field, tenor, and mode” (Martin 1992a: 502). Hence, different registers are to be characterized by different configurations of this three parameters. Figure 2.3 shows the metafunc-

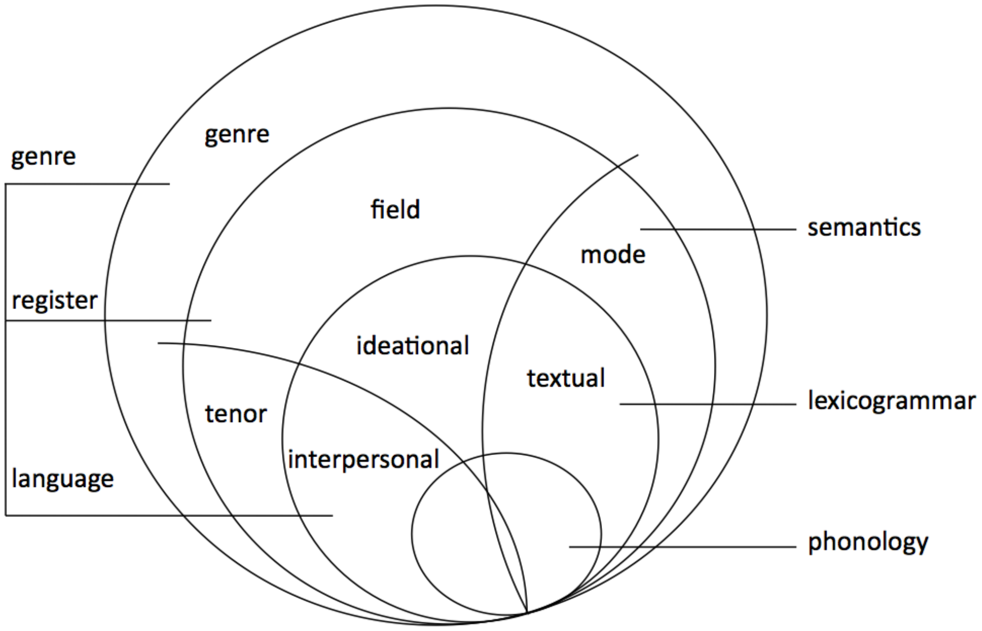


Figure 2.3: Metafunctions in relation to register variables (field, tenor, and mode) and genre (adapted from Martin (1992a, 2007, 2009))

tions in relation to register variables (field, tenor, and mode) and genre. According to SFL, *genre* is a social process. The goals of given text are to be defined in terms of systems of social processes at the level of genre, and the register variables field, tenor, and mode work together to achieve such a text goal.

> For us a genre is a staged, goal-oriented social process. Social because we participate in genres with other people; goal-oriented because we use genres to get things done; staged because it usually takes us a few steps to reach our goals. (Martin & Rose 2007: 8)

Martin (1992a, 1993) argues that since genres are social processes and social processes interact to each other, thereby evolving, a superordinate level to genre in the semantic system is needed: *ideology* (cf. Figure 2.2). "Viewed synoptically, ideology is the system of coding orientations constituting a culture; [...] dynamically it is concerned with the redistribution of power

– of semiotic evolution" (Martin 1992a: 507). The advantages of clearly formulating *genre* as a pattern of *register* patterns in comparison to the previous mentioned approaches, register analysis (cf. Section 2.4) and genre analysis (cf. Section 2.3), are discussed thoroughly in Section 2.5.

The third parameter of SFL's multidimensional model, *instantiation*, refers to the relation between the language system, i.e., the underlying potential of language, and its instances in form of texts (cf. Halliday 2004a: 26, Martin & Rose 2007: 333). The instantiation process itself is determined by the setting of the register variables field, tenor, and mode.

Each single stratum of SFL's language model includes internally three additional organizing parameters, *axiality*, *rank*, and *delicacy*. *Axiality* is concerned with the paradigmatic and syntagmatic ordering in language, i.e., with patterns of choice "in what goes together with what" (Halliday 2004a: 22). *Rank* refers to the units involved in the paradigmatic and syntagmatic axes, e.g., clause, phrase, group, and their associated complexes. Finally, *delicacy* relates to the type-subtype relation organizing paradigmatic axes (cf. Halliday 2004a; Teich 2003).

Section 2.4 presented a brief overview of register analysis, focusing on the theoretical underpinnings of this research. SFL is a sophisticated linguistic model, which makes the analysis of the relations between language and different social contexts possible. It also allows a detailed investigation of discourse variation based on the analysis of concrete linguistic features. The following section sums up the different views on the concepts of *register* and *genre* in the different linguistic approaches presented in Sections 2.3 and 2.4. Finally, Section 2.5 also locates the current objects of study, abstracts and RAs, within the theoretical background of this research, SFL.

2.5 Register & Genre

In the previous sections, three linguistic approaches for the linguistic analysis of discourse were introduced and discussed, i.e., Genre analysis (Section 2.3), Register analysis (Section 2.4) and Systemic Functional Linguistics (Section 2.4.2). Each of these approaches treats the terms *register* and *genre* differently. The first approach, genre analysis, represented mainly by the work of Swales (1990, 2004), only considers the term *genre*. According to this approach, *genre* is defined often very differently, e.g., "a class of communicative events, the members of which share some set of communicative purposes" (Swales 1990: 58) or as "schema" (Hyland 2009: 26). Besides the fact that there is still no general agreement on the definition of *genre* itself

(Hyland 2009; cf. Section 2.3, p. 17), the view of genre as schema has often been criticized, for instance, by Threadgold:

> Genres are not simply schemas or frames for action. They involve, always, characteristic ways of 'text-making' [...], and characteristic sets of interpersonal relationships and meanings.
>
> (Threadgold 1989: 105)

Swales's concept of *genre* has been applied mainly in rhetorical studies, by which research focuses lie on unfolding argumentation structure, and not lexico-grammatical properties of discourse. For these reasons, genre analysis is not the most adequate approach as theoretical background for the intended purposes of this research on abstracts and RAs.

As discussed in the second approach, register analysis (Section 2.5), Biber and his collaborators define the term *genre* "*loosely* [emphasis added] [...] as text categorizations made by the basis of external criteria relating to author/speaker purpose" (Biber 1994: 52) and *register* as "a cover term for any language variety defined in terms of a particular constellation of situational characteristics" (Conrad & Biber 2001: 3). Initially, Biber (1988) exclusively used *genre* rather than *register*. Later, they adopted almost exclusively *register* instead of *genre* (Biber 1995; Biber et al. 1998; Conrad & Biber 2001). Most recently, Biber & Conrad (2009) have included both terms in their work, often considering them interchangeable. Despite this controversy, this approach has been widely used in linguistics for the analysis of language variation, especially due to its theory-looseness and strong reliance on statistical treatment of data. However, the methodology and criteria proposed by Biber and his collaborators for the study of linguistic variation have been criticized as not being ample enough.

> [...] I believe that Biber et al.'s MD theory for register identification is a useful tool in analyzing a register. [...] But that is not enough. [...] Biber et. al.'s proposed criteria for register identification are necessary but not sufficient. If the theory can pool additional linguistic features from the field, the tenor and the mode of discourse, it can establish a more valid profile of each register.
>
> (Ghadessy 2003: 149)

For this reason, their approach does not constitute the theoretical background of this research on abstracts and RAs, although their methods are very inspiring, especially when considering the choice of linguistic features to be evaluated (cf. Section 4.4).

Contrastively, the third approach introduced here, Systemic Functional Linguistics (Section 2.4.2), adopts both terms *register* and, more recently, *genre*, makes a theoretical distinction between them, placing them in two distinctive semiotic levels (Martin 1985, 1997; Matthiessen 2007; cf. Figures 2.2 and 2.3). Martin (1992a: 505-507) argues that considering genre as a pattern of register patterns, just as registers are considered as pattern of linguistic parameters, has several advantages in comparison to other approaches. The first advantage is that placing genre as a further level in the semiotic plane, which is not itself metafunctionally organized, allows the classification of texts cutting across metafunctional components in language. Furthermore, considering genre as a pattern of register patterns accounts for the fact that not in all cultures all combinations of the variables field, tenor, and mode are to be found. Additionally, a theoretical distinction between genre and register facilitates accounting for variety on the sequential unfolding of text as process, and the notion of activity associated with field, tenor, and mode. Finally, Martin addresses the question of genre agnation. He argues that genre is more than just the sum of register parameters:

> The argument here is that social processes are related in ways which complement the valeur determined by looking at them from the perspective of field, mode or tenor alone. Combinations of field, mode and tenor choices in other words enter into relationships with each other which are more than the sum of their parts; to some extent, genres have a life on their own. (Martin 1992a: 507)

In other words, the relationship between register and genre is an interstratal one, with register *realizing* genre, and not a hierarchical one, with one controlling the other (Martin & Rose 2007).

> [...] genre does not determine register variables, any more than register determines linguistic choices. Rather a genre is construed, enacted, presented as a dynamic configuration of field, tenor and mode; which are in turn construed, enacted, presented as unfolding discourse semantic patterns. (Martin & Rose 2007: 309)

Thus, SFL offers a very elaborated model for describing language, allowing a detailed investigation of language variation based on the analysis of concrete linguistic features. For this reason, SFL was chosen as theoretical background of this research.

According to SFL, register variety can be regarded as a continuum of variation throughout all possible settings of the parameters field, tenor, and mode. However, the differences among genres are expected to be more discontinuous and not as easily analyzed along a continuum as register variation. Biber & Conrad summarize the relations between RAs and their individual sections, which they call a case of *embedded genres*, as follows:

> [...] there are cases where genre is embedded in a larger genre. For example, introductory sections in scientific research articles can be analyzed as a genre [...] with its own conventional structure. Form this perspective, the entire introductory section would be regarded as a complete text. These texts represent the genre of "Introduction" because they conform to the expected conventional organization [...]. At the same time, research article introductions are embedded in the larger genre of scientific research article, which has its own conventional structure (e.g., being organized as Abstract, Introduction, Methods, Results, Discussion). (Biber & Conrad 2009: 33)

The main goal of this work is to gain insight into the linguistic characteristics of abstracts in direct comparison with their respective RAs, and to find differences and similarities between them. Hence, finding significant differences regarding the distribution of linguistic features characterizing field, tenor, and mode between abstracts and RAs would imply that they represent different variations of language, at least different *registers*, or conceivably different *genres*.

2.6 Envoi

This chapter has presented and discussed the state-of-the-art linguistic research on abstracts and RAs. First, an overview on the historical development of RAs and abstracts has been provided, thereby naming the most relevant research performed so far. Subsequently, the most frequently adopted approaches to the linguistic analysis of RAs and abstracts has been introduced, followed by a discussion on the choice of the most appropriate approach as theoretical background of this research. Finally, the controversies involving the concepts of genre and register has been addressed in connection with the linguistic model proposed by the theoretical underpinnings of this research.

In order to achieve the goals of this research, linguistic features are to be identified and analyzed that characterize abstracts and RAs, ideally, as different registers. Hence, the concrete design of this research should be rather quantitative than qualitative; the chosen linguistic features are to be extensive enough in order to allow a broad characterization of the texts under study according to field, tenor, and mode of discourse. The analysis of these features should comply with the current methodological approaches in corpus linguistics; and finally, the results are subjected to statistical verification. Hence, Chapter 3 presents and discusses methodological aspects in current linguistic research, setting the scene for the methodology adopted in this research.

Chapter 3

Empirical methods in linguistics

This chapter addresses some issues concerning the methods adopted in current linguistic research. First, a brief overview on empiricism in linguistics is given in Section 3.1. Then, the empirical methods adopted here, from the area of corpus linguistics, are introduced in Section 3.2 together with an exploration of its advantages and disadvantages. Thereafter, Section 3.3 examines the synergies between corpus linguistics, as a methodology, and SFL, as theoretical background for linguistic research on language variation. Finally, Section 3.4 discusses the issue of statistical evaluation of the obtained linguistic data.

3.1 Empiricism in linguistics

Empiricism can be defined as "an approach to a subject (in our case linguistics) which is based upon the analysis of external data (such as texts and corpora[7])" (McEnery & Wilson 2001: 198), in contrast to rationalism, which is based upon introspection, i.e., for the field of linguistics, native speakers of a language who make theoretical claims about this language based on their reflections[8].

[7] The term corpus and its plural, corpora, are defined and discussed in Section 3.2.

[8] Fillmore (1992: 35) describes linguists, who "think" their examples, as follows: "A caricature of the armchair linguist is something like this. He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, 'Wow, what a neat fact!', grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like. (There isn't anybody exactly like this, but there are some approximations)".

The distinction between the empirical and the rational approach to a subject is not a privilege of linguistics, but theoretically exists in any field of scientific research. However, for the discipline of linguistics there is one person responsible for the major discussions involving this natural dichotomy, Avram Noam Chomsky (cf. Chomsky 1957, 1962, 1964, 1965, 1975, 1984, 1988). It is not the aim of this section to describe Chomsky's criticism on empiricism in linguistics thoroughly. Readers interested in this debate find detailed information not only in Chomsky's work itself, but also in several critical reviews, e.g., Haegeman (1991); Horrocks (1987); Matthews (1981). Nevertheless, it is important to mention that, one of Chomsky's major arguments against the use of empirical data in linguistics, i.e., real texts, is that the main goal of linguists should be to explain linguistic competence (i.e., internalized language knowledge) and not to describe and enumerate performance (i.e., externalized utterances) phenomena (McEnery & Wilson 2001: 12). Besides, Chomsky argues that a collection of real texts will never represent the wholeness of language:

> Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so widely skewed that the description [based upon it] would be no more than a mere list. (Chomsky 1962: 159)

The impact of Chomsky's criticism on empirical linguistics lead to the rise of rationalist context-free approaches especially in North America, notably, universal grammar (Chomsky 1965), generative grammar (Chomsky 1965, 1988), transformational grammar (Jackendoff 1974), government and binding (Chomsky 1981), and more recently, minimalist program (Chomsky 1995). Although rationalism dominated the research landscape in linguistics in the 1950s and 1960s, empiricism was never completely abandoned, particularly by Firth (1957, 1968), a leading British linguist, and his followers, e.g., Halliday (1959), and more recently Hoey (2005); Sinclair (1991, 1996, 2003), for whom the central concept in linguistic analysis is the context of situation. Under the influence of these so-called Neo-Firthians, methodological approaches in dealing with naturally occurring language were developed. Within this context, the debate on Chomsky's criticism on empiricism contributed remarkably to the further development of the most acknowledged methodology for empirical research in linguistics, corpus linguistics, which is introduced and discussed in the next section, Section 3.2.

3.2 Corpus linguistics

Until the end of the 1940s, texts under study were mainly "virtual" (Halliday 2004a: 33), i.e., examples thought by grammarians and linguists, with the purpose to illustrate certain categories or descriptions of language. Until then, "real" texts were only available as printed texts. With the development of tape recorders, spoken text became researchable, and with the development of computers, large numbers of texts, both written or spoken, became regularly the object of study in linguistics.

> The study of language is moving into a new era in which the exploitation of modern computers will be at the centre of progress. The machines can be harnessed in order to test our hypotheses, they can show us things that we may not already know and even things which shake our faith quite a bit in established models, and which may cause us to revise our ideas very substantially. In all of this my plea is to trust the text. (Sinclair 1992: 19)

Computers allow linguists to archive and quantitatively analyze real texts on a large scale. Such a collection of texts can be in principle called a *corpus*, which can be defined as "a collection of naturally-occurring language texts, chosen to characterize a state or variety of a language" (Sinclair 1991: 171). However, in current linguistic research, further aspects are to be considered when defining the notion of *corpus*: sampling and representativeness, finite size, machine-readability, and standard reference. For this reason, a more precise definition of *corpus* is given by McEnery & Wilson:

> So a corpus in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration.
> (McEnery & Wilson 2001: 32)

Thus, *corpus linguistics*[9] comprises methodologies allowing the linguistic study of language variety based on authentic texts, i.e., *corpora.* In other

[9] Fillmore (1992: 35) describes corpus linguists as follows: "A caricature of the corpus linguist is something like this. He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is just busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus as the second word of a sentence. (There isn't anybody exactly like this, but there are some approximations)".

words, corpus linguistics is "perhaps best described [...] in simple terms as the study of language based on examples of 'real life' language use". (McEnery & Wilson 2001: 1). However, it is important to reinforce that corpus linguistics is not an area of linguistics, like syntax, semantics, or sociolinguistics, but a group of methodologies for linguistic analysis (Leech 1992: 79). A recent discussion on the status of corpus linguistics as a discipline or method is found in Gries (2010); Teubert (2010a,b). According to Leech (1992: 107) and Biber et al. (1998: 4), the main characteristics of corpus-based linguistic analysis are:

- it is empirical, analyzing the actual patterns of use in natural texts;
- it utilizes a large and principled collection of natural texts, known as a "corpus", as the basis for analysis;
- it makes extensive use of computers for analysis, using both automatic and interactive techniques;
- it depends on both quantitative and qualitative analytical techniques;
- it concentrates on linguistic performance and not on linguistic competence;
- it concentrates on the linguistic description of language instead of linguistic universals;
- it concentrates on the empirical perspective of linguistic analysis and not on the rationalist one.

Chomsky's criticism on corpus linguistics is thereby not completely invalidated (Gilquin & Gries 2009; McEnery & Wilson 2001). Corpus linguists currently ague that the use of corpus does not replace completely the use of intuition (which, however, even for a native speaker can be unreliable, i.e., wrong), rather linguistic research needs corpus *along with* intuition to succeed.

> I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore [...] [but] every corpus I have had the chance to examine, however small, has taught me facts I couldn't imagine finding out any other way. My conclusion is that the two types of linguists [armchair and computer[10]] need another.
>
> (Fillmore 1992: 35)

[10] cf. Footnotes 8 and 9

Linguistic research using corpus linguistics as methodology has considerably increased over the last three decades, mostly due to the improvements in computer techniques in addition to new evidence for recognizing the value of corpora use. Corpus linguistics has been applied in several areas of linguistics, such as lexicography (e.g., Sinclair 2003), grammar (e.g., Biber et al. 1999, 2002), language variation (e.g., Biber 1988, 1990, 1995, 1996, 2006b; Biber & Finegan 1994), translation studies (e.g., Baker 1993, 1995, 1996; Neumann 2003; Teich 2003), among many others.

Corpus analysis can be qualitative or quantitative. Qualitative corpus analysis does not aim to assign frequencies to the linguistic features identified in the corpus. For this kind of analysis, data is only a basis for identification and description of aspects of language use, providing real examples of a given linguistic phenomena under study. In contrast, quantitative corpus researchers identify and classify linguistic features, count them, evaluate them statistically and even develop models to explain what is observed (McEnery & Wilson 2001: 76).

Moreover, corpora can be unannotated or annotated. Unannotated corpora comprise only the plain texts, while annotated corpora are enriched with further linguistic information, e.g., parts-of-speech ("the most basic type of linguistic corpus annotation" (McEnery & Wilson 2001: 46)), morphological information, rhetorical structure information, etc. Initially, linguistic research has been made over unannotated corpora. However, due to the development of linguistic tools for (semi-)automatic annotation of texts, such corpora have gained preference over unannotated ones in linguistic research. The main reason for annotating texts is that through explicit annotation of linguistic information to certain words, texts extracts, etc., this information becomes searchable. The retrieval and therefore quantification and interpretation of linguistic information from annotated texts is the pivotal advantage for using annotated corpora in comparison to unannotated ones.

Although corpus linguistics is primarily adequate for lexical analysis, grammatical studies have been using corpora as object of study very frequently (e.g., Biber et al. 1999). Corpora are very important for grammatical research because of "their potential for the representative quantification of the grammar of a whole language variety" (McEnery & Wilson 2001: 110). Besides, such empirical data can be used for testing hypothesis derived from theoretical linguistic models, as for instance, suggested by Halliday:

> We have always had 'grammars' and 'dictionaries'. [...] at one end are content words, typically very specific in collocation and often of rather low frequency, [...] at the other end are function words, of high frequency an unrestricted collocationally, which relate the content words to each other and enable them to be constructed into various types of functional configuration. (Halliday 1992: 63-64)

> Fundamental work is needed on the probabilistic modeling of systems in a paradigmatic grammar of this kind. But in my view this effort is more likely to be successful if we first find out more of the facts; and that can only be done by interrogating the corpus.
> (Halliday 1992: 76)

Investigation of *actual* language use in large scale is the main advantage of using corpus linguistics as a methodology for studying language structure and variation since other traditional approaches rely on introspection and on exemplarily evidence, based on small samples. However, corpus linguistics is not free of disadvantages. Corpora are always limited in size and only represent a sample of texts collected over a certain period of time. Thus, all obtained results reflect language use only within this time period. Additionally, there is a limitation of language variety covered by a corpus. Generalizing results obtained from corpora are not without hazard since it is impossible to cover all varieties of language in a single corpus.

As mentioned previously, the use of annotated corpora allows the explicit querying of rather implicit linguistic information. Although extremely time consuming, manual annotation of texts delivers a high quality of linguistic information, given that there are precisely formulated annotation guidelines, so that linguists can rely on inter-annotator agreement, avoiding annotation inconsistency. Semi and fully automatic annotation allow the processing of much more texts in much less time. Nevertheless, linguists should be aware of the fact that using such tools for corpora annotation, a given error is always implied, i.e., false annotation, which is innate to the tools in use. In other words, the more linguistic information is automatically added to corpora, the less accurate the annotation will be.

Furthermore, quantitative results obtained using this methodology have to be evaluated for statistical significance, in order to determine how likely the results are due to chance. Although statistical evaluation of data is crucial for helping the linguistic interpretation of results, it is per se not free of limitations (cf. Section 3.4). Hence, corpus linguistics is not the ultimate methodology in current linguistic analysis, but an excelent one, especially as a complement to other approaches.

3.3 Corpus linguistics and SFL

As discussed in Section 2.4.2, Systemic Functional Linguistics (SFL) is a theory of language, while corpus linguistics (CL, Section 3.2) is basically a methodology for analyzing language that can be applied in almost any theoretical framework (Thompson & Hunston 2006: 2). However, there are more synergies between them than one would initially think of. As Halliday points out, there is “a natural affinity between systemic theory and corpus linguistics” (Halliday 2006: 293). Although SFL is a very complex theory, it relies on naturally occurring language, i.e., instances of language in the form of texts (Halliday 2009: 63), and on probabilities (Halliday 2009: 69) for building models for language description. These two issues constitute the the key concerns of CL as well (McEnery & Wilson 2001).

> [...] corpus studies underpin the general principle of functional variation in language; they make it possible to quantify the lexicogrammatical differences among different registers, and to interpret this kind of variation as a redistribution of probabilities.
>
> (Halliday 2006: 294)

Tucker (2006: 102) argues that the incorporation of corpus linguistics into SFL offers additional perspectives on the understanding of social semiotic processes since it illuminates linguistic patterns across the corpus. Additionally, annotation of corpora, whether manual, semi, or fully automated, followed by querying and interpretation of the results provide a further source of linguistic information (cf. Section 3.2), which can be used to support and shape the language model of SFL. However, as stated by Matthiessen (2006: 109), the higher the level to be annotated, the less feasible is automation (cf. Figure 2.2, p. 26). In other words, annotation of word classes can be successfully automated, while the annotation of semantic features according to the SFL model of language becomes a difficult challenge for linguists (Matthiessen 2006: 141). For this reason, SFL studies have only partially applied a corpus linguistic methodology and not many of them have used large corpora so far.

Major future work on the synergies between SFL and CL lies on the development of techniques allowing the annotation and investigation of SFL features on a larger scale. Halliday summarizes the complementariness between SFL and CL as follows:

> A language is a meaning potential, one that is open-ended; the grammatics has to explain how this meaning potential is exploited, and also how it can be enlarged. And this is where I see a complementarity between systemic theory and corpus linguistics. This is not a complementarity of theorising and data-gathering: systemic linguists have always tried to base their descriptions on observable data, while some corpus linguists have proclaimed themselves 'mere data-gatherers' (not without a touch of disingenuousness since I do not think they were really disparaging their own work!), data-gathering is never theory-free, and collecting, managing and interpreting corpus findings is itself a highly theoretical activity.
> (Halliday 2006: 295)

Concerning the quantitative research on the lexico-grammatical properties of abstracts and research articles aimed at here, the decision of adopting SFL as theoretical background has been already addressed on Section 2.4.2. The arguments presented in Sections 3.2 and 3.3 substantiate the presupposition that CL is the most adequate methodology in supporting this research.

> The study of grammatical frequencies is not, I think, some kind of optional extra: such quantitative patterns are a feature of the lexicogrammaticalisation of meaning, the process by which meaning potential becomes effectively without a limit. The project cannot be other than a corpus project. (Halliday 2006: 299)

The core of this research on abstracts and research articles lies in the analysis and interpretation of quantitative linguistic data obtained from the corpus under study. One crucial step in this analysis is the statistical evaluation of the results, which aims to clarify how likely these results are due to chance. Section 3.4 thus discusses the main issues in statistical evaluation of linguistic data based on current practices of corpus linguistics.

3.4 Statistical evaluation of linguistic data

As mentioned in Section 3.2, corpus linguistics is an empirical methodology allowing scientific quantitative research in linguistics. After quantitative data is obtained, linguists are to describe and interpret the results carefully in order to explain given linguistic phenomena. The first step in treating data is to describe them as accurately and revealingly as possible. Thereafter, linguists have to evaluate and interpret the data, most likely through the testing of previously formulated hypotheses. The knowledge of statistical methods is imperative for linguists to achieve these aims.

Statistical methods can be generally divided into *descriptive* statistics and *analytical* statistics. Descriptive statistics provides linguists with methods for optimal description and visualization of data. Analytical statistics offers linguists adequate methods for significance testing of data as a means to hypothesis-testing. This section does not aim to provide a comprehensive overview on statistics for linguistics, but only addresses the relevant methods and techniques applied in this research, complying with current practice of corpus linguistics. The formulae of the statistical techniques used in this research are not described here since they are standard in current statistics. Comprehensive work in statistics for linguistics is found in e.g., Baayen (2008); Gries (2008b, 2009a,b); Manning & Schütze (1999); Oakes (1998); Rasinger (2008). Thus, Section 3.4.1 presents the techniques for data visualization and description applied here. The following section, 3.4.2, discusses the techniques for evaluation of data significance used in this research. Finally, Section 3.4.3 addresses the issue of limitations of statistics in quantitative linguistic analysis in general.

The open-source software *R*[11], which is not only a language but also an environment for statistical computing and powerful graphics, has been chosen as the tool for loading, processing, visualizing, and statistically analyzing quantitative data in this research. The most comprehensive review of R and its applications is currently given by Crawley (2007). Dedicated work on the application of R in linguistics is found in e.g., Baayen (2008); Gries (2008b, 2009a,b). The R-codes used in this study are not described in this section, but presented in Appendix A.6.

3.4.1 Descriptive statistics

Descriptive statistics helps linguists describe and graphically display quantitative data. Frequencies of occurrence of linguistic features, which are the most typical kind of quantitative data in corpus linguistics, are usually displayed in form of tables, scatterplots, bar or pie charts.

There is an extremely useful way of displaying data using R, which is called *boxplot with notches*. Figure 3.1 displays a graph of an example of fictitious data. Boxplots show the distribution of data, either a single sample or several samples simultaneously, in a box around the horizontal thick line,

[11] `http://www.r-project.org/` (accessed: 18 July 2010).

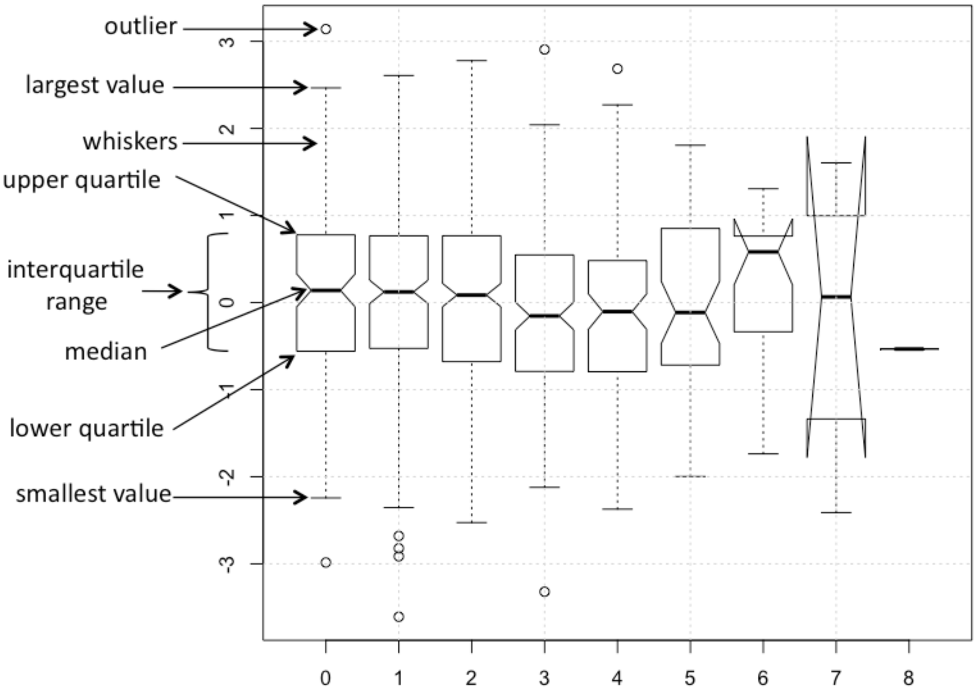


Figure 3.1: Fictitious example of a *boxplot with notches*

the *median*[12]. The distribution of data is divided into four equal *quartiles*. The top of the box shows the upper quartile and the bottom of the box indicates the lower quartile. The vertical dashed lines are called *whiskers*. They show either the maximum (and minimum, respectively) value or 1.5 times the interquartile range of data, whichever is smaller (Crawley 2007: 155). *Outliers* are values which are more than 1.5 times the interquartile range above the third quartile or below the first quartile. They are always plotted individually. In other words, in case of not having outliers, whiskers always indicate the maximum and minimum values. The use of *median* instead of the well known (arithmetic) *mean* is more advantageous since medians are less sensitive to outliers, which can distort the mean (Neumann 2008: 83). Boxplots are not only useful for showing the location and spread of data, but also for indicating asymmetry of data, i.e., *skewness*, in the

[12] According to Baayen (2008: 21ff.), "the median is obtained by ordering the observations from small to large, and then taking the central value (or the average of the two central values when the number of observations is even".

sizes of the upper and lower part of the boxes. For instance, in Figure 3.1, sample 1 has a symmetrical distribution of data around the median, while sample 5 shows a much higher range of data above its median than below. Furthermore, the *notches*, drawn as a waist around the median, aim to give an idea about the significance of the differences between several medians. According to Crawley (2007: 157), "boxes in which the notches do not overlap are likely to prove to have significantly different medians under an appropriate test". Thus, such graphical illustration of data is highly helpful to gain first insights in linguistic data. For example, Figure 3.1 indicates that there is probably no significant difference between the medians of samples 0, 1, and 2, while the comparison between samples 1 and 3 will probably show significant different medians. Moreover, samples 6 and 7 are examples of either samples with small sizes and/or with high variance. In such cases, the notches are not displayed as usual, but fold downwards/upwards. This is how R warns users about possible invalidity of the notch test. However, linguists must bear in mind that each of these "*prima facie*" evidence of significant difference between medians *must* be statistically tested properly (Gries 2009a: 205). The proper tests for this purpose are discussed in Section 3.4.2.

Besides boxplots, *histograms* are used also in visualizing data distribution within a single sample in the present study. Histograms are excellent for showing the *mode*, i.e., the value occurring most frequently, the spread, and the skew of a set of data (Crawley 2007: 162). Through such graphs, linguists can easily visualize whether data are normally distributed or not. The concept of normal distribution of data is of central importance in analytical statistics, particularly concerning the choice of statistical tests for significance to be applied to data (cf. Section 3.4.2).

Figure 3.2 illustrates an example of a histogram for a fictitious feature x, showing, that in this case, data are *not* normally distributed. Normally distributed data are characterized by symmetrical distribution of data around the mode and equal values for mean, median, and mode. Boxplots also provide information concerning the tendency to normality of data since the thick line representing the median would be precisely in the middle of the box and the whiskers would be equally high to the top and the bottom of the median in case of normally distributed data (cf. Figure 3.1). Nevertheless, the results provided by histograms are more accurate.

Thus, descriptive statistics provides linguists not only with relevant information about the characteristics of experimental data, but also with convenient techniques for visualizing such data. However, in order to eval-

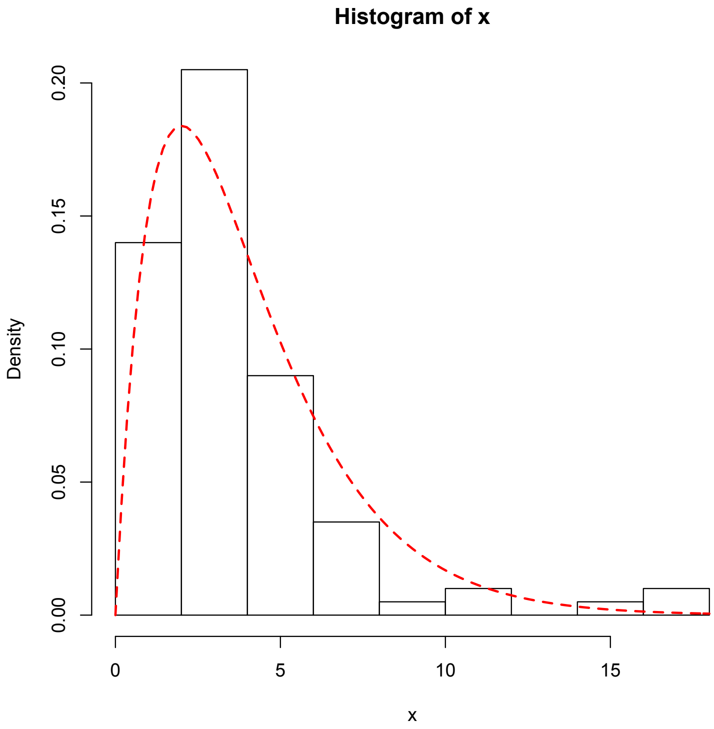


Figure 3.2: Fictitious example of a *histogram*

uate and interpret data in more detail, e.g., to determine whether values between different samples are statistically significant, linguists need specific statistical tests, which are part of analytical statistics.

3.4.2 Analytical statistics

Analytical statistics, also known as *inferential* statistics, is used to evaluate and interpret sample data. They are especially used in the domain of hypothesis and significance testing (i.e., *deductive* approach to research), which is of pivotal importance in a quantitative corpus linguistic methodology. Generally, the hypotheses which are to be tested with the help of quantitative data and analytical statistics, should be formulated *prior* to the collection of data (Gries 2009b: 13). According to Bortz & Döring (2006: 7), a hypothesis to be tested should be a statement concerned with a single phenomenon. It should have the structure of a conditional sentence and be potentially falsifiable. The linguists' own hypothesis is called *alternative hypothesis* and it is denoted as H_1. However, analytical statistics does not allow the verification of the trueness of H_1. Analytical statistics provide linguists with methodologies and tests for showing that the logical counterpart of the researcher's own hypothesis is most probable *not* true. The logical counterpart of H_1 is called the *null hypothesis* and is denoted as H_0.

> The most essential part of the statistical approach of hypothesis testing is that, contrary to what you might expect, you do *not* try to prove that your alternative hypothesis is correct – you try to show that the null hypothesis is most likely(!) not true, and since the null hypothesis is the logical counterpart of your alternative hypothesis, this in turn lends credence to the alternative hypothesis. In other words, in most cases you wish to be able to show that the null hypothesis can *not* account for the data so that you can adopt the alternative hypothesis.
>
> (Gries 2009a: 183ff.)

Therefore, the null hypothesis should state that variables are normally or randomly distributed; or that there is no difference between groups, samples, or variables; and that in case of existing differences, these are due to chance (Gries 2009b: 13).

A ficticious example of hypothesis formulation in case a linguist is interested in the relation between gender and slang use could be:

H_1: There is a relationship between the gender of the speaker and the use of slang in language

H_0: There is no relationship between the gender of the speaker and the use of slang in language

The data obtained from the analysis of an adequate corpus would be then tested for the probable negation of H_0.

The choice of the proper tests to be applied to hypothesis-testing, however, is very dependent on the kind of variables one is dealing with. It is not the aim of this section to present all possible combinations of tests and kinds of variables. This section addresses only the tests used in this research. Readers find more detailed discussion on this issue in e.g., Gries (2008b); Rasinger (2008). The process of choosing the adequate test for hypothesis-testing should take the following questions into consideration (Crawley 2007; Gries 2009a):

- Are the observations of the samples independent of each other?
- Are there outliers in the data?
- Are the values normally distributed?
- Are the variances homogeneous?

Correlation of the samples, outliers, non-normality, and heterogeneous variances can invalidate inferences made by some standard tests. One of the most frequently used standard statistical test for sample comparison in linguistics is the *t-test*. However, the t-test may *only* be used if the observations of the samples are independent; the data is normally distributed; and the variances of the samples are homogeneous. Therefore, *before* applying the t-test for hypothesis-testing, linguists should answer the four questions mentioned above.

For this research on abstracts and research articles, there is no relationship between the observations of the samples since data are randomly extracted from a corpus. Besides, the corpus itself is a random selection of texts of the domains under study (cf. Section 4.1). However, as will be discussed in Chapter 5, there are some outliers. The existence of outliers is detected using boxplots with notches for data visualization (cf. Section 3.4.1).

The condition of normal distribution is tested using not only histograms, but most importantly using the *Shapiro-Wilk* test. The Shapiro-Wilk test computes the value for the parameter *W*. For Shapiro-Wilk test, the hypotheses to be tested are:

$\mathbf{H}_1$: The data deviate from a normal distribution; W $\neq$ 1

$\mathbf{H}_0$: The data do not deviate from a normal distribution; W = 1

If the values of W are smaller than 1 and the corresponding *p-values* (which estimate the probability that a result could have occurred by chance) are smaller than 0.05, then, there is a probability of at least 95% that the results are not due to chance and therefore statistically significant. This being the case, it means that H_0 has to be rejected and consequently the data is not normally distributed (cf. Gries (2009a: 208) and Crawley (2007: 282)). Most of the data obtained in this research is *not* normally distributed (cf. Chapter 5). This observation corroborates Gries's statement that "natural linguistic data are only rarely normally distributed" (Gries 2009a: 210).

Finally, the homogeneity of variances is tested using the *Fligner-Killeen* test since most of the data in this research do not allow the use of the usual test for this purpose, the *F-test*, which also requires normally distributed data. Additionally, the *Fligner-Killeen* test has the advantage of being not sensitive to outliers (Crawley 2007: 293). The results of this test often indicate non-homogenous variances for the data in this study (cf. Chapter 5).

Since the data in this research often did not conform[13] to the prerequisites for using the t-test for significance testing, the *Wilcoxon rank-sum* test is the test chosen for determining whether differences between samples are statistically significant or not. The Wilcoxon rank-sum test, also known as U-test, is the non-parametric alternative to the t-test. It has no prerequisites to be tested and can be used for non-normally distributed data. Again, if the calculated *p-values are less than 0.5*, the null hypothesis can be rejected and the compared samples are significantly different from each other. One advantage of this test is that it is said to be *more conservative* than the t-test (in case it could have been applied to the data). In other words, if the Wilcoxon rank-sum test computes a significant difference, “it would have been even more significant under a t-test” (Crawley 2007: 298).

Another statistical test used in this research is the *chi-square* test (χ^2). The chi-square test is one of the most important tests used very frequently in linguistics. It tests the association degrees between categorial, i.e., nominal, variables and it can be used for determining whether such samples are significantly different from each other. Basically, the chi-square test compares the observed values with the appropriate set of expected ones. In other words, it compares whether “our expectations and our actual data correspond” (Rasinger 2008: 145). The requirements in using the chi-square test are all observations are independent of each other; 80% of the expected frequencies are larger or equal to 5; and all expected frequencies are larger than 1 (Gries 2009b: 152). If the calculated p-value is smaller than 0.05 the samples are significantly different from each other.

However, the chi-square test implies an intrinsic approximation inherent to its internal calculation formula, i.e., the so called expected values. For this reason, the *Fisher's exact test* should always be preferred over the chi-square test. The Fisher's exact test allows the exact calculation of the significance of the deviation from a null hypothesis, rather than relying on an approximation, as in the case of the chi-square test. The only inconvenience of the Fisher's exact test is that it can not be performed ”by hand” due to the complexity of its formula. However, R has a function for calcu-

[13] According to Gries (2009a: 241), the t-test function implemented in R does not require homogeneity of variances. However, there is still the violation of normal distribution, which is the knock-out criteria for not using the t-test.

lating the Fisher's exact test, which nevertheless may require lots of CPU time and even cause overload. Alternatively, the function for calculating the Fisher's exact test can be set for simulation instead of exact calculation. Thus, when treating contingency tables, first a Fisher's exact test is to be performed. In case it does not work because of overload, then a chi-square test is performed. If the chi-square test returns a warning message that the results may be not correct, then a Fisher-test with simulation is performed.

The last sort of statistical methods to be introduced here is the so called *multivariate statistics*. Multivariate statistics comprises a class of statistical methods which is fundamentally different from the previously presented statistical tests. This is because it does not look for variation in variables, but for "structure in the data" Crawley (2007: 731). For this purpose, two different multivariate statistical techniques are applied in this study: *hierarchical agglomerative cluster analysis* and *principal component analysis*.

In contrast to all previously mentioned statistical tests, *hierarchical agglomerative cluster analysis* requires *no* hypotheses formulation prior to data testing (i.e., *inductive* approach to research). It is, therefore, a *data-driven* exploratory approach, by which all the steps "do not involve any (potentially biased) human decisions" (Gries 2006: 129).

> The idea behind hierarchical cluster analysis is to show which of a (potentially large) set of samples are most similar to one another, and to group these similar samples in the limb of a tree. Groups of samples that are distinctly different are placed in other limbs. The trick is in defining what we mean by 'most similar'.
>
> (Crawley 2007: 742)

The two parameters for determining group sampling are the similarity measure and the amalgamation rule. The similarity measure is a correlational measure, calculated based on the cosine distance between the vectors in the matrix of numerical data. The amalgamation rule calculates for every possible amalgamation "the sums of squared differences from the mean of the potential cluster, and then the clustering with the smallest sum of squared deviations is chosen" (Gries 2009b: 317). Detailed information on the calculation steps behind hierarchical agglomerative cluster analysis is found in Baayen (2008); Crawley (2007); Gries (2009b).

Figure 3.3 shows a fictitious example of a *cluster dendrogram*, which is the output of a hierarchical agglomerative cluster analysis. In this fictitious example, USA states are clustered based on data concerning the number of

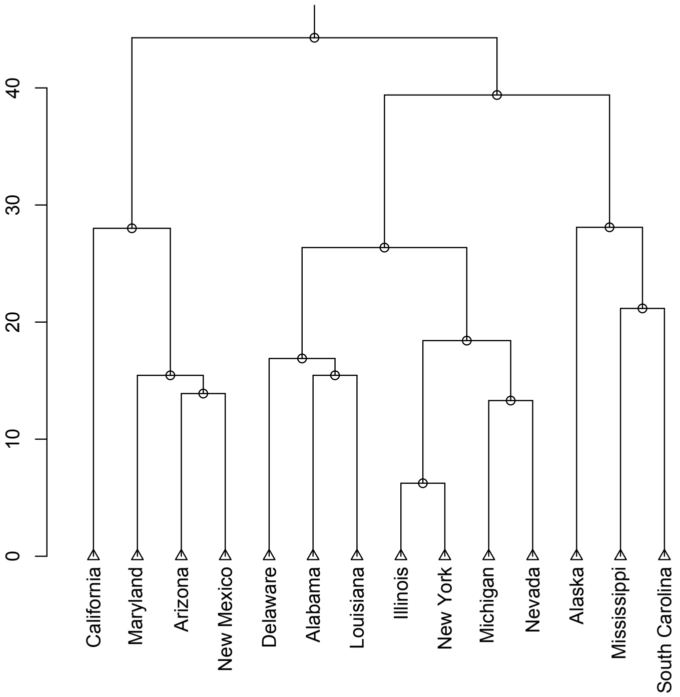


Figure 3.3: Fictitious example of a *cluster dendrogram* (adapted from R Documentation on General Tree Structures)[14]

arrested people in these states. According to Figure 3.3, the first cluster is composed of California, Maryland, Arizona and New Mexico, by which the two last states having a more similar profile of arrests when compared to the first two ones. This first cluster is also very distinct in comparison to the fourth cluster, comprising the states of Alaska, Mississippi and South Carolina. This figure provides the researcher interested in the profiling of arrests in the USA with grouping of states in degrees of similarities. The next step would thus be the interpretation of this clustering to be performed by the researcher. For each dendrogram, every single sample is plotted as its own cluster at the bottom of the dendrogram. Starting from every single sample, a vertical line is drawn upwards reflecting the degree of similarity

[14] `http://sekhon.berkeley.edu/stats/html/dendrogram.html` (accessed: 03 November 2010).

of these samples based on its height. Finally, the vertical lines are grouped together by horizontal lines. The longer the vertical lines, the more distinct the clusters are. Hierarchical agglomerative cluster analysis thus provides researchers with an objective categorization of data, which can be certainly seen as a valuable component of the data evaluation and interpretation process.

Principal component analysis (PCA), the second multivariate statistical method applied in this study, aims to "find a small number of linear combinations of the variables so as to capture most of the variation in the dataframe as a whole" (Crawley 2007: 731). When analyzing n variables, e.g., linguistic features, the researcher is dealing with n vectors in a dataframe matrix and therefore with a n-dimensional space. It is quite comfortable for the researcher to handle the data, if the whole dataframe is reduced to a small number of combinations of the original data explaining the variance in the data matrix. Ideally, this number of components is reduced to two or three, so that a n-dimensional space can be reduced, analyzed and visualized in a bi- or tridimensional space. Furthermore, the components found in this analysis are very useful in finding patterns of correlation between the data and the features investigated. The concrete application of this method in this study and the results obtained are discussed thoroughly in Section 5.2.2. A detailed discussion on PCA and its applications can be found in Jolliffe (2002); Kline (1994).

Although *analysis of variance* (ANOVA) has become very popular in linguistic research lately, it was not used in this study. The reasons for this are twofold. First, ANOVA is a parametric technique like the t-test, which can be used to analyze variability within- and between groups. Therefore, similarly to the t-test, ANOVA requires normally distributed data among other prerequisites. As mentioned before, the data in this study does not comply with this condition very often. Second, there are other statistical techniques, as discussed previously, which deliver relevant results to support linguists' interpretation of data.

> [...] [T]he fact that t-tests and ANOVAs are the parametric techniques to investigate between-groups and within-groups variance does not prove that other statistical methods cannot also yield interesting results. It is for these reasons that, in spite of their appeal at a superficial glance, t-tests and ANOVAs do not enjoy a central status here. (Gries 2006: 145)

Statistical methods are fundamental parts of quantitative linguistic studies. They help in describing and visualizing data and are essential for the process of evaluation and interpretation of large amount of quantitative data. For this reason, statistics plays a very important role in the overall corpus linguistics methodology. However, as already partially discussed, statistical methods are not free of constraint. The next section, Section 3.4.3, deals with the limitations of statistical techniques.

3.4.3 Limitations of statistics

As discussed in Sections 3.4.1 and 3.4.2, statistics is fundamental in the process of description, visualization, evaluation and interpretation of quantitative results in linguistics. As any other methodology in science, it has not only advantages, but also limitations.

Statistics is only suitable in the study of quantitative phenomenon, not for qualitative studies. Consequently, statistics is not adequate for the study of individual items; it deals only with groups as an aggregate of items. For this reason, several statistical techniques are not adequate for small samples or groups of items.

Knowing that every corpus represents a given language only partially, studies on larger corpora probably deliver better results in comparison to smaller corpora. However, no matter how large the corpus under study is, it will never represent the wholeness of language. Thus, the results of a corpus analysis primarily represent the properties of the studied corpus. Inferences and generalizations beyond the corpus under study should therefore be formulated with caution.

Another important aspect is that all statistical techniques are based on probabilities and approximations. Hence, statistical results do not represent the absolute trueness of facts, but only likeness and tendencies. For this reason, it is very important that the collection, analysis and interpretation of data is performed carefully, otherwise the statistical results may be misleading.

Finally, applying statistical techniques to data requires good statistical knowledge of the researchers. They should clearly understand not only the constraints in applying a certain test in advance, but also what is behind the formulae and prerequisites in order to avoid misused tests and wrong interpretations of the results. It does not matter whether statistics is used for a deductive or inductive approach to research, the most important step in linguistic research is still the careful interpretation of data by the researcher.

The previous chapters provided information about the state-of-the-art of linguistic research on abstracts and research articles followed by the theoretical background and aims of this research (cf. Chapter 2) as well as the empirical corpus linguistic methodology and statistical evaluation of obtained data (cf. Chapter 3). Chapter 4 discusses the concrete design of this research, its corpus, the hypotheses to be tested, and the linguistic features chosen for the quantitative analysis.

Chapter **4**

Research design

This chapter presents the design of this research, which aims to investigate linguistic differences between abstracts and their research articles (RAs) across several scientific domains. Section 4.1 introduces the corpus and its design, followed by the description of its processing and annotation steps in Section 4.2. The hypotheses tested in the empirical analysis are then formulated in Section 4.3. Finally, Section 4.4 deals with the choice of the linguistic features for the empirical analysis.

4.1 Corpus

This research is conducted in the context of the project "Linguistic profiles of interdisciplinary registers[15]" (*Linguistische Profile interdisziplinärer Register*; henceforth LingPro; Teich & Holtz (2009); Teich & Fankhauser (2010)). LingPro investigates the linguistic genesis of interdisciplinary registers at the boundaries of computer science with some other established scientific discipline, and aims to develop register profiles of the texts produced in selected scientific "cross"- disciplines (bioinformatics, computational linguistics, computational engineering and microelectronics). For these purposes, a corpus of journal articles from several scientific disciplines was built (Darmstadt Scientific Text Corpus; henceforth DaSciTex). DaSciTex covers nine scientific domains, and has a three-way partition: computer science; "mixed" disciplines, i.e., interdisciplinary domains involving computer science and one "pure"-discipline (computational linguistics, bioinformatics, computer-aided design, and microelectronics); and "pure" disciplines, i.e.,

[15] `http://www.linglit.tu-darmstadt.de/index.php?id=lingpro_projekt` (accessed: 21 July 2010).

Discipline	Journals	Year
Computer science	1. Journal of Algorithms	2004-2006
	2. Journal of Computer and System Science	2005-2007
Linguistics	1. Language	2003-2006
	2. J. of Linguistics	2006 (42:1)
	3. Functions of Language	2005-2006
	4. Linguistic Inquiry	2005-2006
Biology	1. Gene	2004-2006
	2. Nucleic Acid Research	2006
Mechanical engineering	1. Chemical Engineering and Processing	2006-2007
	2. Chemical Engineering Science	2006(10)-(1)2007
	3. International J. of Heat and Mass Transfer	2006(10)-(1)2007

Table 4.1: Overview of the sources of the AbstRA corpus

disciplines from which the mergers with computer science are built (linguistics, biology, mechanical engineering, and electrical engineering). The choice of the journals included in DaSciTex[16] was based on experts' recommendations from the chosen disciplines.

The corpus used in the present research on abstracts and RAs is a subset of DaSciTex. It is called AbstRA (Abstracts and Research Articles Corpus), hereafter. It comprises only the disciplines of *computer science*, *linguistics*, *biology*, and *mechanical engineering*. The reason for deciding to work only with these four disciplines is twofold. First, a selection for a subcorpus had to be made due to time constraints since this research aims to quantitatively analyze several linguistic features in a given time span. Second, it is expected that in case of existing domain specific differences between abstracts and their RAs, such differences are more distinctive between disciplines that are expected not to be very similar to each other. Additionally, the four chosen disciplines represent one discipline from the

[16] Readers find detailed information about DaSciTex on the project's homepage (cf. footnote 15).

humanities (*linguistics*), one discipline from the natural sciences (*biology*), one discipline from engineering (*mechanical engineering*), and finally the last one, *computer science*. It is quite different from the previously mentioned disciplines, being perhaps more similar to mathematics.

Table 4.1 shows the journals from which the articles were extracted, and the time span when the articles were collected. The focus of the journal selection criteria lies on the representativity of the journals of the "mixed" disciplines and then on the acknowledged journals for the corresponding "pure" discipline. This was the rationale for the choice of journals to represent a given discipline. For instance, for the discipline of *biology*, the journals chosen were *Gene* and *Nucleic Acid Research*. From LingPro's point of view which aims to study the development of the register of *bioinformatics*, these are certainly the two most adequate journals in the DaSciTex corpus. The same is valid for all other domains. Since AbstRA, the corpus under study here, is taken from the DaSciTex corpus, the constraints for building up the corpus apply. Hence, the AbstRA corpus *does not* aim to be a representative of the entirity of each of these four disciplines. Yet, the AbstRA corpus is adequate in representing language patterns of both abstracts and RAs in the chosen domains. According to Biber et al., the issues on corpus design can be summarized as follows:

> [...] [I]t is important to be realistic. Given constraint on time, finances, and availability of texts, compromises often have to be made. Every corpus will have limitations, but a well-designed corpus will still be useful for investigating a variety of linguistic issues.
>
> (Biber et al. 1998: 250)

The number of texts and *tokens* per discipline and for the whole AbstRA corpus is shown in Table 4.2. Tokens in this case represent the number of running words. All 94 texts were obtained from online journals mostly in original PDF[17]-format. This format, however, does not allow any further annotation and querying of linguistic information of texts. For this reason, all texts of the corpus were converted to HTML format using the AnnoLab suite[18](Eckart 2006; Eckart & Teich 2007). UTF-8[19] encoding was used to assure that as many as possible of the original characters remained intact.

[17] Adobe Portable Document Format; `http://www.adobe.com/products/acrobat/adobepdf.html` (accessed: 21 July 2010).

[18] `http://www.annolab.org/` (accessed: 21 July 2010).

[19] UTF-8 is defined by the Unicode Standard [UNICODE]. Descriptions and formulae can also be found in Annex D of ISO/IEC 10646-1 [ISO.10646]; `http://www.iso.org/` (accessed: 24 July 2010).

Discipline	Abstracts		Research articles*	
	Texts	Tokens	Texts	Tokens
Computer science	27	4,772	27	134,890
Linguistics	14	2,565	14	126,442
Biology	24	7,428	24	80,295
Mechanical engineering	29	4,386	29	79,398
	94	19,151	94	421,025

*Research articles not including their abstracts

Table 4.2: Overall size of the AbstRA corpus

After this conversion into HTML, each of the subcorpus of RAs comprised around 150,000 tokens per discipline. The number of texts was chosen, so that the number of tokens would be similar over the four disciplines. However, the resulting texts are not completely clean (e.g., erroneous splitting / contraction of tokens). For some types of linguistic investigations, this quality of data may be acceptable, but for this present study, it is crucial to have them absolutely clean. For this reason, all 94 texts were manually cleaned, although this procedure is very time consuming. Regrettably, the texts of the disciplines *biology* and *mechanical engineering* were more "dirty" than the others. This was also due to the use of many tables and figures, which were converted into gibberish and were therefore discarded. This is the reason for the variation in the final number of tokens per discipline in the RA subcorpus (cf. Table 4.2). The number of tokens in the abstract subcorpus is dependent on the RA subcorpus since abstracts are part of RAs. Hence, the AbstRA corpus contains only manually cleaned texts consisting of a total of 19,151 tokens of abstracts texts and 421,025 tokens of RAs texts. This number of tokens, both for abstracts and RAs complies with the size criteria recommended by Biber et al. (1998: 248) for *grammatical* inquiry. According to Biber et al. (1998: 249), lexicographic studies however demand very large corpora.

Although the size of AbstRA is *not* adequate for linguistic research concerning lexical features, quantitative results in this area are still relevant in gaining insights into the lexical characteristics of abstracts and their RAs.

Due to copyright restrictions for the texts included in DaSciTex and consequently in AbstRA, both corpora cannot be made freely available.

4.2 Corpus processing and annotation

The AbstRA corpus is composed of several annotation layers together with metadata information for each of its texts. The metadata for bibliographical information is based on the TEI[20] standard. Additionally, metadata information for the situational parameters of field, tenor, and mode of discourse is provided for all texts in AbstRA. The tenor of discourse is invariably *expert-to-expert* and the mode of discourse is *written-to-be-read* since all the texts are journal articles (cf. Halliday & Hasan 1989). The field of discourse varies according to the disciplines. It is encoded in the metadata information as the keywords provided by each RA. The management of this data is achieved via JabRef[21], an open source bibliography reference manager. Figure 4.1 shows the typical metadata information for a text in the corpus with field, tenor, and mode information. Each of the `{text}`-fields in Figure 4.1 contains the corresponding information for each single text in AbstRA, thus allowing the tracking of bibliographical and context information for all texts in the corpus.

All processing steps of AbstRA are managed also by the AnnoLab suite (Eckart 2006; Eckart & Teich 2007). AnnoLab is a modular extensible framework, which is able to deal with texts annotated at multiple levels of linguistic organization (multi-layer annotations). Each layer is represented in an XML document and the different layers are connected to the text data via stand-off references. AnnoLab is written in Java 1.5 and can use Apache UIMA[22] to manage linguistic processing chains. Data are stored in an eXist[23] native XML database, which can be queried with XQuery, an XML query language[24].

With the help of AnnoLab, a processing pipeline was built for tokenization, part-of-speech (PoS) tagging, lemmatization, and syntactic parsing of the texts in AbstRA (cf. Figure 4.2). The tagger incorporated in AnnoLab is the *TreeTagger*, a language independent part-of-speech tagger (Schmid 1994a,b). TreeTagger's English parameter file was trained on the PENN

[20] `http://www.tei-c.org/index.xml` (accessed: 24 July 2010).

[21] `http://jabref.sourceforge.net/` (accessed: 24 July 2010). The native file format used by JabRef is BibTex, the standard LaTeX bibliography format.

[22] `http://uima.apache.org/` (accessed: 25 July 2010).

[23] `http://exist.sourceforge.net/` (accessed: 25 July 2010).

[24] `http://www.w3.org/TR/xquery/` (accessed: 25 July 2010).

```
@ARTICLE{ text-year,
  author = { text },
  title = { text },
  journal = { text },
  year = { text },
  volume = { text },
  pages = { text },
  number = { text },
  month = { text },
  abstract = { text },
  keywords = { text },
  owner = { text },
  pdf = {archive\A\text-year.pdf},
  field = { text },
  tenor = {expert-to-expert},
  mode = {written-to-be-read},
  timestamp = { text },
  url = { text }
}
```

Figure 4.1: Metadata annotation of the AbstRA corpus

Treebank and its tag set[25] (Marcus et al. 1993). According to Schmid (1994b), the TreeTagger achieves 96.36% accuracy. The syntactic parser incorporated in AnnoLab is the statistical *Stanford Parser*[26], which is a program for determining the grammatical structure of sentences in form of e.g., nominal and prepositional phrases (Klein & Manning 2003a,b). Its performance is 86.36% (Klein & Manning 2003a: 423), which is excellent for a statistical syntactic parser. Figure 4.2 shows the processing pipeline used for AbstRA.

The metadata information and linguistic annotations are stored separately in different layers, one for each type of annotation. Appendix A.1 illustrates such a multi-layer annotation in XML. The annotated corpus can be queried over strings, annotations of a single layer and multiple layers. However, retrieving information from the corpus requires the use and combination of several query tools, depending on what is being queried. The

[25] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-Tagset.pdf` (accessed: 25 July 2010).

[26] `http://nlp.stanford.edu/software/lex-parser.shtml` (accessed: 25 July 2010).

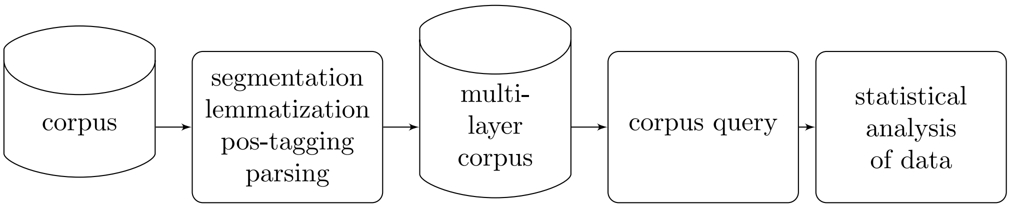


Figure 4.2: AbstRA corpus processing pipeline

IMS Corpus Workbench (IMS-CWB; Christ 1994) is employed for querying the PoS-layer. IMS-CWB is a set of tools for the manipulation of large, linguistically annotated text corpora, which includes the IMS Corpus Query Processor (CQP; Christ et al. 1999), a specialized search engine for linguistic research. The query over the parsed layer is performed using XQuery[27], an XML query language, over the AnnoLab database. Finally, data analysis was performed using either Microsoft Excel®, WordSmithTools 5.0 (Scott 2008) or R, depending on type of analysis.

Sections 4.1 and 4.2 described the rationale for building the AbstRA corpus as well as its processing and annotation steps. The next section approaches the formulation of hypotheses to be tested empirically in this study.

4.3 Hypotheses

As mentioned in Section 1.1, the main goal of this thesis, which is also the *research problem* here, is to gain insight into linguistic characteristics of abstracts in direct comparison with their respective RAs, in the expectation of finding significant differences between them. For the purpose of quantitative investigation of authentic language used in abstracts and RAs, the AbstRA corpus was built (cf. Section 4.1). The main *variables* in this corpus are *text type*[28] (abstracts / RAs) and *domain* (computer science / linguistics / biology / mechanical engineering). It is possible to formulate one *hypothesis*, which is to be tested by the empirical analysis, for each of these two variables. Moreover, a third hypothesis is formulated to investigate a possible register/genre variation between abstracts and RAs. *Regis-*

[27] `http://www.w3.org/TR/xquery/` (accessed: 25 July 2010).

[28] The term text type is used here in a broad, linguistic, non-technical sense, just as an equivalent to the German term *Textsorte*.

ter variation is associated with *functional* variation of language, reflected in variation of the linguistic choices within different registers. Contrastively, *genre* variation is associated with the social processes and their linguistic instances in language use to fulfill communicative *purposes* (Biber & Conrad 2009; Martin 1992a). Such variation directly results in different patterns of linguistic features, which can be quantitatively investigated and statistically evaluated. The next sections present and discuss the hypotheses to be tested in this research.

4.3.1 Variation according to text type

Abstracts and their RAs are intuitively different. This research aims to find significant differences between these two text types. Such differences are to be detected at both the *lexical* and the *grammatical* level through the quantitative analysis of linguistic features. The concrete set of linguistic features to be investigated in this study is discussed thoroughly in Section 4.4. The null ($H1_0$) and the alternative ($H1_1$) hypotheses for Hypothesis 1 are thus formulated as follows:

$\mathbf{H1}_1$: The quantitative analysis of linguistic features reveals statistically significant differences between abstracts and their RAs at both lexical and grammatical levels.

$\mathbf{H1}_0$: The quantitative analysis of linguistic features reveals *no* statistically significant differences between abstracts and their RAs at both lexical and grammatical levels.

4.3.2 Variation according to domain

Although possible differences between abstracts and RAs exist, domain specific differences within each of these two text types are also presumed. In other words, differences within abstracts of the four domains under study are expected to be found. The same is valid for RAs. This is because different disciplines tend to express their knowledge in different ways (cf. Halliday & Martin 1993). Such differences are to be detected at both the *lexical* and the *grammatical* level through the quantitative analysis of linguistic features. Again, the set of linguistic features to be investigated is presented and discussed in Section 4.4. Therefore, the null ($H2_0$) and the alternative ($H2_1$) hypotheses for Hypothesis 2 are formulated as follows:

$\mathbf{H2}_1$: The quantitative analysis of linguistic features reveals statistically significant differences across domains for abstracts and RAs at both lexical and grammatical levels.

H2$_0$: The quantitative analysis of linguistic features reveals *no* statistically significant differences across domains for abstracts and RAs at both lexical and grammatical levels.

4.3.3 Variation according to the context of situation

The last issue to be addressed here is the theoretical question whether abstracts and their RAs are distinct registers or maybe even distinct genres (cf. Section 2.5). The answer to this question is directly dependent on results showing different patterns of linguistic features. For registers, different patterns reflect differences in the configuration of the parameters of the context of situation, i.e., field, tenor and mode of discourse. According to Martin & Rose (2007: 309), differences in genres would reflect "the field, tenor and mode selections that genres do and do not share". While register variation can be seen as a continuum, genre variation is more discrete (Biber & Conrad 2009: 33).

Nevertheless, register/genre differences are undoubtedly reflected in differences in the configuration of the parameters of the context of situation and differences in the realizations of concrete linguistic features. Thus, such differences are to be detected at both the lexical and the grammatical level through the quantitative analysis of linguistic features.

As for the other two hypotheses, the set of linguistic features to be investigated is discussed in Section 4.4. The null (H3$_0$) and the alternative (H3$_1$) hypotheses for Hypothesis 3 are therefore formulated as follows:

H3$_1$: Abstracts and their RAs show different configurations of the parameters of context of situation field, tenor, and mode of discourse.

H3$_0$: Abstracts and their RAs *do not* show different configurations of the parameters of context of situation field, tenor, and mode of discourse.

These three hypotheses are to be statically tested in the empirical analysis of the AbstRA corpus in Chapter 5 and subjected to corroboration or refutation in Chapter 6. In order to test these hypotheses, adequate observable linguistic *features* must be selected, as it is not possible to entirely test all possible features in a language. However, such linguistic features should not be randomly selected. They should be deduced from *indicators*, which, in turn, are adequate for the *characterization* of different language variations since the aim of this research is to show that abstracts and RAs differ linguistically. The criteria for the selection of such indicators and their corresponding linguistic features are discussed in the next section.

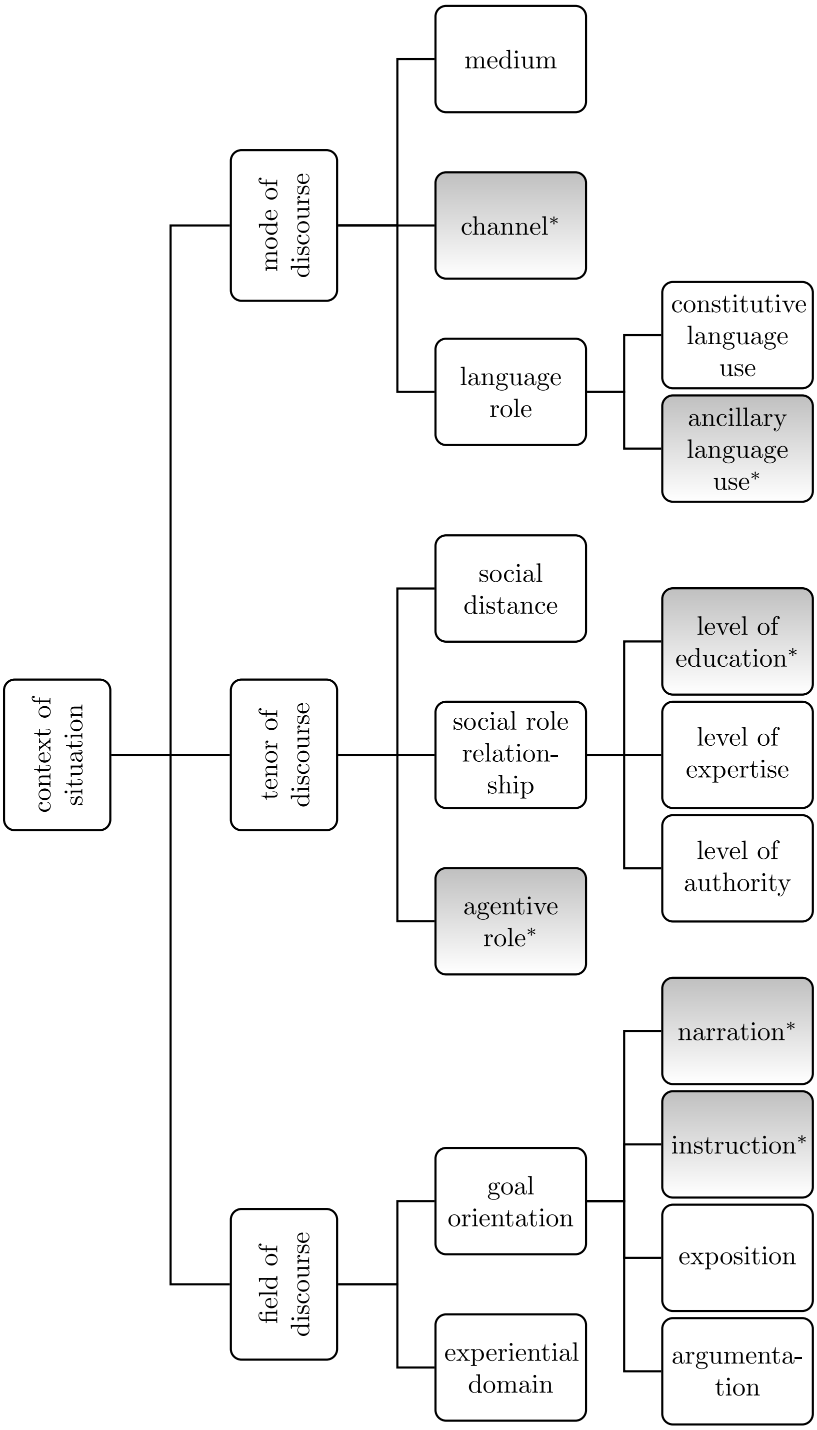
4.4 Indicators for empirical analysis

According to the underlying theoretical framework of this research, Systemic Functional Linguistics (SFL), language and context are "inextricably linked" (Thompson 2004: 10). SFL's language model is composed of three metafunctions, i.e., ideational, interpersonal, and textual, whose configurations are determined by the situational context in which language is being used. These three metafunctions are linguistically expressed through the three parameters of the context of situation, field, tenor, and mode of discourse, respectively. According to SFL, language variation therefore unavoidably reflects a variation in the configurations of field, tenor, and mode (cf. Figure 2.3, p. 27). For this reason, the parameters field, tenor, and mode of discourse are adequate *parameters* for the study of language variation (cf. Halliday & Hasan 1989; Neumann 2003, 2008; Steiner 1983; Teich 2003).

Knowing that these parameters are linguistically realized at the lexicogrammatical level, different configurations of these reflect directly into different realizations of very concrete linguistic *features*, which are adequate *indicators* for quantitative empirical analysis. The selection of these features is a crucial step on the research design. The criteria for feature selection used in this research is based on previous works on register variation (e.g., Biber 1988, 1995, 2006c; Biber & Finegan 1994; Biber et al. 2007; Neumann 2003, 2008). They also take the design and size of the AbstRA corpus into consideration, which are not adequate for applying any kinds of features, as discussed in Section 4.1.

According to Halliday & Hasan (1989: 26), the field of discourse is concerned with "what's going on", the tenor of discourse covers "who is taking part", and the mode of discourse deals with the "role assigned to language" in the context of situation where language is functioning. These parameters are further classified into several subcategories, which are going to be briefly described here, followed by a discussion of their applicability to this research, as well as of the typical linguistic features that may be used to characterize the AbstRA corpus according to these parameters. Figure 4.3 summarizes the classification of the parameters of context of situation and the forthcoming discussion.

The field of discourse, the first parameter of the context of situation, can be subdivided into the parameters *experiential domain* and *goal orientation.* The *experiential domain* is also called "the nature of social activity" (Halliday & Hasan 1989), which deals with the topic of the context where



*= parameter not addressed in this research

Figure 4.3: Parameters of context of situation

language is being used. Vocabulary (e.g., most frequent words, special terminology), distribution of lexical words (nouns, lexical verbs, adjectives, adverbs), and keywords[29] are therefore suitable linguistic features, i.e., indicators, for investigating variation in the experiential domain. The *goal orientation* is concerned with the purpose of the text and is further classified into argumentation, exposition, instruction, and narration. This definition of *goal orientation* is *very similar* to the *genre* definition according to Martin (1992a). Hence, differences in the parameter of *goal orientation* very much likely reflect *genre variation*.

Since the AbstRA corpus contains only scientific papers, the expected goal orientation types are argumentation and/or exposition. A typical feature used as an indicator for the characterization of argumentative texts is the use of modals, e.g., prediction, necessity, and possibility modals, for the argumentative discourse is "designed to persuade the addresse" (Biber 1988: 111). Prediction modals, i.e., *will*, *would*, *shall*, are directly related to whether a given event *will* happen or not. Necessity modals, i.e., *ought*, *should*, *must*, are directly concerned with the obligation and necessity of events happening and are therefore directly *persuasive* (Biber 1988: 150). Finally, possibility modals, i.e., *can, may, might, could*, deal with the possibility of an event occurring. Possibility modals are mainly used to express uncertainty or lack of precision concerning the information presented, or to discuss different perspectives on a given topic (Biber 1988: 106, 150). For this reason, necessity and possibility modals are of pivotal importance in characterizing argumentative discourse.

Past tense is a linguistic feature that serves as an indicator of expository texts since a low frequency of occurrence of past tense is expected for event-oriented, static, descriptive, or expository discourse (Biber 1988: 109). Additionally, the features high frequency of nouns, high type/token ratio, and the presence of nominalizations and passives are further indicators of expository texts since they are characterized by a highly informational discourse (Biber 1988: 104).

The second parameter of the context of situation, tenor of discourse, can be further divided into three types: *agentive role*, *social role relationship*, and *social distance*. *Agentive role* deals with the identification of information-giver and information-taker in language. Taking into consideration that the AbstRA corpus contains only written texts, further

[29] "A word which is positively key occurs more often than would be expected by chance in comparison with the reference corpus" (`http://www.lexically.net/downloads/version5/HTML/index.html`) (accessed: 29 July 2010).

investigation of this parameter is pointless in this research since for all texts the information-givers are the authors and the information-takers are the readers of the texts. *Social role relationship* deals with the hierarchy, i.e., relationship of power between participants, i.e., information-givers and information-takers, in the situation, and is sub-classified into *level of authority*, *level of expertise*, and *level of education.*

A typical linguistic feature characterizing *level of authority* is the use of modality. This is specially valid for necessity modals, i.e., *ought*, *should*, *must* since they allow information-givers, in this case the authors, to achieve a more distant and powerful social role in the communication process. The *level of expertise* is linguistically realized in the form of terminology, use of keywords and nominalizations, and long sentences. The *level of education* is identifiable, for instance, in teaching environments, where communication between people of different levels of education take place, e.g., students, who are not proficient in language and therefore make grammatical mistakes. Knowing that the texts in the AbstRA corpus are published in journals, i.e., from *expert-to-expert*, it is unnecessary to investigate this parameter in this research. *Social distance* is concerned with the style of the communication taking place, i.e., level of formality. This parameter is similar to Biber's dimension "involved vs informational production". Typical linguistic features characterizing this dimension are, for instance, high frequency of nouns, word length, the use of prepositional phrases, high type/token ratio (Biber 1988).

Finally, mode of discourse, the third parameter of the context of situation, is sub-classified into *language-role*, *channel*, and *medium. Language role*, i.e., the kind of language being used, can be further divided into ancillary and constitutive language use. Ancillary language, i.e., language supporting a nonverbal action, is typical of spoken language use. It is characterized, for instance, by the high frequency of ellipsis. This parameter is not relevant in this research since the AbstRA corpus contains only written texts. Constitutive language use is characterized by the full use of the language potential. Typical linguistic features characterizing constitutive language use are indicative mood, declarative sentences, and high density of lexical words. *Channel* deals with the physical conditions of the communication (Halliday & Hasan 1989). In all the texts of the AbstRA corpus its configuration is "graphic" since all texts were published in print. For this reason, this parameter is not going to be addressed in this research. Finally, *medium* covers the differences between spoken and written language, and their special cases like speeches, which although written, are expected to be heard by the audience. Although all texts in AbstRA are of the kind “written-to-be-read”, differences between texts could potentially still occur (Neumann 2008: 65). Typical linguistic features characterizing this parameter are high density of lexical words, and grammatical complexity[30], among others.

This research aims to characterize abstracts and RAs through a quantitative study of several of the above discussed linguistic features over the AbstRA corpus, in order to test the hypotheses formulated in Section 4.3. The choice of the features considered also other issues, like for instance, the feasibility of their quantification according to the methodology to be adopted. The selected features can be grouped into three categories: shallow features, i.e., those reflecting general characteristics of the corpus, lexical, and grammatical features. The linguistic features chosen for the empirical analysis are

- Shallow features
    - Sentence length (i.e., words/sentence)
    - Type/token ratio
    - Distribution of lexical words
- Lexical features
    - Lexical density
    - Distribution of the most frequent lexical items
    - Keywords: keyness of the most frequent lexical items[31]
- Grammatical features
    - Distribution of modals
    - Distribution of passives
    - Distribution of nominalizations
    - Grammatical complexity (as grammatical phrases, e.g., NPs, PPs, PPs embedded in NPs, etc. . .)

Table 4.3 summarizes the relationship between the *parameters* field, tenor and mode of discourse, and the *indicators*, i.e., the linguistic features chosen for the quantitative analysis, taking into account that one feature may be *indicative* of more than one *parameter* subcategory simultaneously.

[30] For detailed information, see Section 5.1.3.4.

[31] Keyness: how much a keyword is a keyword; always in comparison to reference corpora (for details cf. Section 5.1.2.3).

Parameter	Parameter subcategory (what the features are indicative of)	Indicator (linguistic feature)
Field of discourse	Experiential domain	Keywords Lexical words
	Goal orientation	Lexical words Modals Nominalizations Passives Type/token ratio Pronominalization
Tenor of discourse	Social role relationship	Keywords Modals Nominalizations Sentence length
	Social distance	Lexical words Type/token ratio Grammatical complexity
Mode of discourse	Language role	Lexical density
	Medium	Grammatical complexity Lexical density

Table 4.3: Relationship between *parameters* and *indicators*

This research follows a twofold empirical analysis plan. First, a deductive empirical analysis is to be performed, by which the selected features are to be quantitatively determined and statistically evaluated for significance and hypothesis testing. Then, an inductive empirical analysis is to be performed. The purpose of the inductive empirical analysis is to corroborate (or not) the results of the deductive empirical analysis. Since inductive analysis formulates no hypothesis to be tested prior to the analysis, the obtained results in such a "theory-free" analysis are good indicators of how adequate the hypothesis and features chosen for the deductive empirical analysis are.

The present chapter initially introduced AbstRA, the corpus under study, its design, processing steps, and annotations. Then, hypotheses to be tested by the empirical analysis were formulated, followed by a discussion on adequate indicators, i.e., features, to empirically test these hypotheses. The next chapter, Chapter 5, presents the results of both the deductive and inductive empirical analysis and their evaluation. The relationship between the actual results of the empirical analysis, the indicators, and the parameters is then discussed in Chapter 6.

Chapter 5

Empirical analysis

This chapter presents the results of the deductive and inductive empirical analyses and their evaluation over the AbstRA corpus. Section 5.1 presents the results of the deductive empirical analysis, comprising the data for shallow (Section 5.1.1), lexical (Section 5.1.2), and grammatical analysis (Section 5.1.3). Then, Section 5.2 discusses the results of the inductive empirical analysis including the hierarchical agglomerative cluster analysis (Section 5.2.1) and principal component analysis (Section 5.2.2).

The correlation between the empirical results presented here and the parameters for context of situation is discussed in Chapter 6. When applicable, statistical evaluation of data is discussed as well as the procedural method used in R, which is described in detail in Appendix A.6. Generally, results concerning abstracts and RAs are presented first, followed by the respective discussion on the domain variation in each sub-corpus. For the purpose of comparison with other established corpora, the following reference corpora are used in this research, depending on the feature being considered: the Freiburg-LOB Corpus of British English[32] (FLOB), the Freiburg-Brown corpus of American English[33] (Frown), and the British National Corpus[34] (BNC).

[32] `http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/index.html` (accessed: 08 August 2010).

[33] `http://www.helsinki.fi/varieng/CoRD/corpora/FROWN/` (accessed: 08 August 2010).

[34] `http://www.natcorp.ox.ac.uk/` (accessed: 08 August 2010).

5.1 Deductive empirical analysis

In this section, the results of the deductive empirical analysis are presented and discussed. They comprise shallow (Section 5.1.1), lexical (Section 5.1.2), and grammatical (Section 5.1.3) features analysis. Since a deductive analysis is used, null hypotheses prior to the analyses themselves are formulated. Based on the obtained results and on statistical significance testing, the formulated null hypotheses are then refuted or not.

5.1.1 Shallow features

This section discusses the results for the shallow features, starting from the more general to the more specific feature. The first results to be discussed concern the distribution of parts-of-speech (Section 5.1.1.1), followed by the values for words per sentence (Section 5.1.1.2). Section 5.1.1.3 introduces the concept of type/token ratio and presents the corresponding data for the AbstRA corpus. The last of the shallow features, distribution of lexical words, is then discussed in Section 5.1.1.4.

5.1.1.1 Parts-of-speech

Data concerning the distribution of parts-of-speech are the first entrance to quantitative analysis of annotated corpora, like the AbstRA corpus. They constitute the basic data about the corpus and provide an overview on the frequencies of word classes indicating potential interesting linguistic phenomena for further study. As mentioned in Section 4.2, AbstRA is tagged for part-of-speech according to the tagset used by the TreeTagger[35]. The meaning of each tag in this tagset is found in Appendix A.2. Only the most interesting findings are discussed here. Therefore, not all parts-of-speech are going to be addressed in this section. Table 5.1 (cf. p. 72) presents the results of the distribution of parts-of-speech, as raw frequencies of occurrence and as their respective percentages, for abstracts and RAs. Some interesting observations emerge from the data. The data discussed

[35] The tagset used by the TreeTagger is a refinement of the Penn Treebank tagset (cf. Appendix A.2). Therefore, the TreeTagger includes additional information in its tags, as follows:

> The second letter of the verb part-of-speech tags is used to distinguish between forms of the verb "to be" (B), the verb "to have" (H), and all the other verbs (V). So, "VHD" is the POS tag for the past tense form of the verb "to have", i.e. for the word "had". `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/` (accessed: 31 July 2010).

in this section is highlighted in bold in Table 5.1. The other reason for depicting such an amount of data in tables throughout this thesis is to account for replicability of all quantitative analytical steps carried out in the course of this study.

The first observation derived from the data in Table 5.1 concerns the distribution of modals. Modals (MD) occur less than half as frequently in abstracts (0.33%) as in RAs (0.77%). This observation is already a corroboration that the choice of having modals within the set of features for the quantitative analysis may lead to relevant results.

Although the percentage of singular (NN) and plural nouns (NNS) is very similar for both sub-corpora, proper nouns in singular and plural forms (NP + NPS) occur twice more frequently in abstracts (10.05%) than in RAs (5.22%). This may be an indication that abstracts show a wider vocabulary variety than RAs. Such an assumption has to be checked for in the quantitative analysis of nouns distribution. However, the distribution of adjectives (JJ, JJR, JJS) is very similar for both sub-corpora. This observation is quite striking since adjective modify nouns; the more frequent nouns are, the more frequent adjectives are usually expected to be. This could be an indication that abstracts tend to present information objectively, not qualifying or modifying the nouns involved.

The frequency of occurrence of adverbs (RB + RBR + RBS) is lower in abstracts (2.25%) than in RAs (3.39%). Again, this could be an indication that abstracts tend to present information more precisely and in a more concise manner than RAs, not qualifying or modifying the verbs involved.

Finally, present tense (V*P + V*Z) occurs less often in abstracts (2.52%) than in RAs (3.36%), while past participle (V*N) occurs slightly more often in abstracts (3.33%) than in RAs (3.03%). This indicates that the use of verb tense may be, as assumed formerly, a proper indicator of discourse variation between abstracts and RAs.

Such observations are a good primary indication for the adequacy of the chosen linguistic features for differentiating between abstracts and RAs, e.g., distribution of lexical words, modals, passives. In other words, they are very promising for detecting potential significant differences between abstracts and RAs, which are investigated in detail in the following sections. At this point, no further statistical evaluation of these data is performed.

PoS-tag	Abstracts		Research articles	
	F	%	F	%
#	0	0.00	60	0.01
$	0	0.00	3	0.00
”	24	0.13	45	0.01
(	206	1.08	7,409	1.76
)	207	1.08	7,399	1.76
,	1,800	9.40	18,195	4.32
:	476	2.49	7,246	1.72
“	11	0.06	159	0.04
CC	456	2.38	10,508	2.50
CD	357	1.86	13,733	3.26
DT	1,624	8.48	41,715	9.91
EX	11	0.06	609	0.14
FW	3	0.02	93	0.02
IN	1,893	9.88	48,369	11.49
JJ	**1,419**	**7.41**	**30,080**	**7.14**
JJR	**44**	**0.23**	**1,140**	**0.27**
JJS	**34**	**0.18**	**883**	**0.21**
LS	24	0.13	2,221	0.53
MD	**64**	**0.33**	**3,255**	**0.77**
NN	**3,342**	**17.45**	**70,006**	**16.63**
NNS	**1,052**	**5.49**	**22,142**	**5.26**
NP	**1,924**	**10.05**	**21,941**	**5.21**
NPS	**0**	**0.00**	**46**	**0.01**
PDT	19	0.10	300	0.07
POS	23	0.12	27	0.01
PP	179	0.93	5,674	1.35
PP$	44	0.23	1,498	0.36
RB	**411**	**2.15**	**13,500**	**3.21**
RBR	**14**	**0.07**	**566**	**0.13**
RBS	**6**	**0.03**	**209**	**0.05**
RP	7	0.04	435	0.10
SENT	686	3.58	24,865	5.91
SYM	349	1.82	3,856	0.92
TO	286	1.49	7,027	1.67
UH	0	0.00	107	0.03
VB	48	0.25	2,515	0.60
VBD	101	0.53	2,431	0.58
VBG	4	0.02	127	0.03
VBN	**34**	**0.18**	**511**	**0.12**
VBP	**102**	**0.53**	**2,480**	**0.59**
VBZ	**212**	**1.11**	**6,138**	**1.46**
VH	3	0.02	281	0.07
VHD	5	0.03	164	0.04
VHG	3	0.02	92	0.02
VHN	**35**	**0.18**	**2**	**0.00**
VHP	**27**	**0.14**	**762**	**0.18**
VHZ	**0**	**0.00**	**871**	**0.21**
VV	232	1.21	6,908	1.64
VVD	95	0.50	2,415	0.57
VVG	281	1.47	5,968	1.42
VVN	**568**	**2.97**	**12,231**	**2.91**
VVP	**127**	**0.66**	**2,902**	**0.69**
VVZ	**142**	**0.74**	**4,750**	**1.13**
WDT	82	0.43	2,510	0.60
WP	4	0.02	374	0.09
WP$	3	0.02	60	0.01
WRB	48	0.25	1,212	0.29
Σ	19,151	100.00	421,025	100.00

Table 5.1: Distribution of parts-of-speech for the AbstRA corpus

The distribution of parts-of-speech for abstracts across the four different disciplines, i.e., computer science, linguistics, biology, and mechanical engineering, is presented in Table 5.2. Interestingly, abstracts from biology use the lowest number of modals (0.22%), which is almost half of the frequency of occurrence of modals in the discipline with their highest amount, i.e., linguistics (0.43%), followed by computer science (0.42%) and mechanical engineering (0.39%). This is already an indication of domain specific variation on the use of modals in abstracts, which is going to be investigated thoroughly in Section 5.1.3.1. Contrastively, in biology abstracts nouns (37.99%) are much more frequent than in the other disciplines, i.e., mechanical engineering (32.74%), computer science (28.37%), and linguistics (27.53%), specially proper nouns with 20.42%. Such initial observations indicate, that abstracts from the discipline of biology may significantly differ from abstracts from the other disciplines.

Finally, Table 5.3 shows the results for the distribution of parts-of-speech for RAs across the four disciplines. Similar to abstracts, the sub-corpus of RAs also shows that biology is the discipline with the lowest frequency of modals (0.46%), being again almost half of the values encountered for the other disciplines: computer science (0.93%), linguistics (0.81%), and mechanical engineering (0.75%). There is also a notable difference in the distribution of nouns across disciplines in the sub-corpus of RAs. Once more, biology is the discipline with the highest number of nouns (31.30%), followed by mechanical engineering (29.11%), computer science (25.80%), and linguistics (24.55%). In comparison to the results for the abstract sub-corpora, however, RAs show a lower frequency of proper nouns for biology (9.03%). However, RAs from the discipline of biology still show the highest frequency of proper nouns in comparison to computer science (4.55%), linguistics (4.36%), and mechanical engineering (3.82%).

The analysis of this first shallow feature, distribution of parts-of-speech, allows one to gain a first insight into the characteristics of the AbstRA corpus. Differences between abstracts and RAs, and across disciplines were observed. The first observations corroborate the initial assumptions that there are linguistic differences between abstracts and their RAs. Additionally, they support the choice of the linguistic features to be investigated further. Thus there are possible significant differences to be found. In case there are no differences found in the distribution of parts-of-speech at all, it would probably be worthless to continue with this research. Since this is not the case, Section 5.1.1.2 discusses the results for the next shallow feature, the distribution of words per sentence.

PoS-tag	Computer science		Linguistics		Biology		Mechanical engineering	
	F	%	F	%	F	%	F	%
”	7	0.15	5	0.19	11	0.15	1	0.02
(	84	1.76	31	1.21	68	0.92	23	0.52
)	85	1.78	31	1.21	68	0.92	23	0.52
,	175	3.67	109	4.25	1387	18.67	129	2.94
:	156	3.27	56	2.18	185	2.49	79	1.80
“	3	0.06	1	0.04	2	0.03	5	0.11
CC	103	2.16	77	3.00	150	2.02	126	2.87
CD	118	2.47	28	1.09	124	1.67	87	1.98
DT	451	9.45	254	9.90	416	5.60	503	11.47
EX	4	0.08	5	0.19	1	0.01	1	0.02
FW	1	0.02	1	0.04	0	0.00	1	0.02
IN	462	9.68	323	12.59	552	7.43	556	12.68
JJ	344	7.21	250	9.75	448	6.03	377	8.60
JJR	10	0.21	11	0.43	12	0.16	11	0.25
JJS	24	0.50	2	0.08	6	0.08	2	0.05
LS	14	0.29	0	0.00	5	0.07	5	0.11
MD	**20**	**0.42**	**11**	**0.43**	**16**	**0.22**	**17**	**0.39**
NN	**885**	**18.55**	**432**	**16.84**	**928**	**12.49**	**1,097**	**25.01**
NNS	**265**	**5.55**	**171**	**6.67**	**377**	**5.08**	**239**	**5.45**
NP	**204**	**4.27**	**103**	**4.02**	**1,517**	**20.42**	**100**	**2.28**
PDT	8	0.17	4	0.16	2	0.03	5	0.11
POS	12	0.25	6	0.23	3	0.04	2	0.05
PP	75	1.57	37	1.44	47	0.63	20	0.46
PP$	26	0.54	5	0.19	8	0.11	5	0.11
RB	111	2.33	81	3.16	130	1.75	89	2.03
RBR	1	0.02	5	0.19	6	0.08	2	0.05
RBS	4	0.08	1	0.04	1	0.01	0	0.00
RP	1	0.02	1	0.04	4	0.05	1	0.02
SENT	203	4.25	97	3.78	203	2.73	183	4.17
SYM	209	4.38	14	0.55	52	0.70	74	1.69
TO	90	1.89	64	2.50	76	1.02	56	1.28
VB	20	0.42	8	0.31	9	0.12	11	0.25
VBD	4	0.08	6	0.23	47	0.63	44	1.00
VBG	3	0.06	1	0.04	0	0.00	0	0.00
VBN	11	0.23	1	0.04	9	0.12	13	0.30
VBP	31	0.65	18	0.70	21	0.28	32	0.73
VBZ	60	1.26	40	1.56	48	0.65	64	1.46
VH	1	0.02	0	0.00	1	0.01	1	0.02
VHD	0	0.00	0	0.00	4	0.05	1	0.02
VHG	0	0.00	2	0.08	0	0.00	1	0.02
VHP	12	0.25	2	0.08	13	0.18	8	0.18
VHZ	7	0.15	2	0.08	9	0.12	9	0.21
VV	78	1.63	52	2.03	64	0.86	38	0.87
VVD	10	0.21	12	0.47	49	0.66	24	0.55
VVG	81	1.70	36	1.40	81	1.09	83	1.89
VVN	139	2.91	76	2.96	172	2.32	181	4.13
VVP	54	1.13	29	1.13	38	0.51	6	0.14
VVZ	52	1.09	32	1.25	29	0.39	29	0.66
WDT	30	0.63	16	0.62	20	0.27	16	0.36
WP	1	0.02	3	0.12	0	0.00	0	0.00
WP$	1	0.02	1	0.04	1	0.01	0	0.00
WRB	22	0.46	12	0.47	8	0.11	6	0.14
Σ	4,772	100.00	2,565	100.00	7,428	100.00	4,386	100.00

Table 5.2: Distribution of parts-of-speech for abstracts

PoS-tag	Computer science		Linguistics		Biology		Mechanical engineering	
	F	%	F	%	F	%	F	%
#	3	0.00	4	0.00	53	0.07	0	0.00
$	0	0.00	3	0.00	0	0.00	0	0.00
”	22	0.02	13	0.01	8	0.01	2	0.00
(	2,027	1.50	2,170	1.72	1,978	2.46	1,234	1.55
)	2,019	1.50	2,169	1.72	1,978	2.46	1,233	1.55
,	5,956	4.42	5,829	4.61	3,486	4.34	2,924	3.68
:	1,914	1.42	2,539	2.01	1,980	2.47	813	1.02
“	86	0.06	49	0.04	10	0.01	14	0.02
CC	2,705	2.01	3,418	2.70	2,175	2.71	2,210	2.78
CD	3,458	2.56	3,339	2.64	3,934	4.90	3,002	3.78
DT	14,565	10.80	12,491	9.88	6,102	7.60	8,557	10.78
EX	271	0.20	241	0.19	42	0.05	55	0.07
FW	25	0.02	47	0.04	6	0.01	15	0.02
IN	15,444	11.45	14,936	11.81	8,580	10.69	9,409	11.85
JJ	8,712	6.46	10,140	8.02	5,456	6.79	5,772	7.27
JJR	320	0.24	371	0.29	168	0.21	281	0.35
JJS	606	0.45	136	0.11	100	0.12	41	0.05
LS	794	0.59	426	0.34	399	0.50	602	0.76
MD	**1,261**	**0.93**	**1,029**	**0.81**	**371**	**0.46**	**594**	**0.75**
NN	**22,459**	**16.65**	**18,306**	**14.48**	**13,125**	**16.35**	**16,116**	**20.30**
NNS	**6,208**	**4.60**	**7,217**	**5.71**	**4,756**	**5.92**	**3,961**	**4.99**
NP	**6,142**	**4.55**	**5,511**	**4.36**	**7,252**	**9.03**	**3,036**	**3.82**
NPS	8	0.01	23	0.02	9	0.01	6	0.01
PDT	157	0.12	62	0.05	35	0.04	46	0.06
POS	7	0.01	2	0.00	18	0.02	0	0.00
PP	2,665	1.98	2,024	1.60	505	0.63	480	0.60
PP$	551	0.41	620	0.49	196	0.24	131	0.16
RB	4,572	3.39	4,919	3.89	2,095	2.61	1,914	2.41
RBR	156	0.12	230	0.18	99	0.12	81	0.10
RBS	100	0.07	60	0.05	23	0.03	26	0.03
RP	138	0.10	188	0.15	49	0.06	60	0.08
SENT	8,842	6.55	7,275	5.75	4,184	5.21	4,564	5.75
SYM	1,618	1.20	527	0.42	565	0.70	1,146	1.44
TO	2,264	1.68	2,430	1.92	1,148	1.43	1,185	1.49
UH	7	0.01	96	0.08	2	0.00	2	0.00
VB	1,015	0.75	728	0.58	249	0.31	523	0.66
VBD	100	0.07	650	0.51	1,035	1.29	646	0.81
VBG	27	0.02	66	0.05	15	0.02	19	0.02
VBN	112	0.08	147	0.12	135	0.17	117	0.15
VBP	796	0.59	853	0.67	352	0.44	479	0.60
VBZ	2,574	1.91	1,841	1.46	578	0.72	1,145	1.44
VH	102	0.08	128	0.10	34	0.04	17	0.02
VHD	6	0.00	86	0.07	44	0.05	28	0.04
VHG	27	0.02	46	0.04	9	0.01	10	0.01
VHN	0	0.00	1	0.00	0	0.00	1	0.00
VHP	289	0.21	219	0.17	154	0.19	100	0.13
VHZ	354	0.26	255	0.20	127	0.16	135	0.17
VV	3,075	2.28	2,164	1.71	783	0.98	886	1.12
VVD	375	0.28	963	0.76	692	0.86	385	0.48
VVG	1,870	1.39	1,710	1.35	1,186	1.48	1,202	1.51
VVN	3,231	2.40	3,397	2.69	2,772	3.45	2,831	3.57
VVP	1,427	1.06	934	0.74	312	0.39	229	0.29
VVZ	2,115	1.57	1,569	1.24	482	0.60	584	0.74
WDT	768	0.57	1,090	0.86	320	0.40	332	0.42
WP	49	0.04	308	0.24	12	0.01	5	0.01
WP$	19	0.01	28	0.02	9	0.01	4	0.01
WRB	477	0.35	419	0.33	108	0.13	208	0.26
Σ	134,890	100.00	126,442	100.00	80,295	100.00	79,398	100.00

Table 5.3: Distribution of parts-of-speech for research articles

5.1.1.2 Sentence length

Sentence length is a feature marking structural complexity and elaborateness in discourse (Biber 1988: 47, Biber & Conrad 2009: 152). Since abstracts are supposed to summarize the information presented in RAs in a compact form (cf. Section 2.2.2), it is expected that abstracts have longer sentences, i.e., *higher sentence length*, in comparison to their RAs. Example 5.1 from a mechanical engineering abstract and Example 5.2 from a RA from computer science illustrate this assumption.

(5.1) Heat transfer to an immersed sphere from fluidized uncoated sand particles of different mean size and size distribution is compared with that from coated sand particles of equal size extracted from two full-scale fluidized bed boilers for different superficial gas velocities and mean particle diameters from 350 to 646 [mu]m. [abstract.C3.3]

(5.2) We first observe that for this specific problem, a much simpler algorithm achieves the same 2-approximation. [RA.A.20]

Thus, the null hypothesis to be tested, H_0, and its counterpart, the alternative hypothesis H_1, can be formulated as follows:

H_1: Abstracts have significantly higher sentence length in comparison to their RAs.

H_0: Abstracts *do not* have significantly higher sentence length in comparison to their RAs.

Sentence length is quantitatively determined by calculating the ratio between the number of words and the number of sentences for every single text in the AbstRA corpus. This calculation is performed by WordSmith Tools. The values for sentence length for each single text of the AbstRA corpus is found in Table 5.4. One abstract in biology shows unexpected high sentence length (278; marked within a rectangle), being composed of just one single sentence with 278 words. This value has been manually checked and its corresponding text proved to be a unique exception in the whole AbstRA corpus. In order to avoid such an outlier, this text and its corresponding RA (29.8108; also marked within a rectangle) are not taken into consideration in the statistical evaluation of data. Therefore, they were removed from the corpus and from the data set.

As discussed in Section 3.4.1, such data is better visualized through the plot *boxplot with notches*. This plot is generated by R from the data in Table 5.4, which are previously saved in two separate tables, one for all abstracts

Abstracts				Research articles			
Computer science	Linguistics	Biology	Mechanical engineering	Computer science	Linguistics	Biology	Mechanical engineering
22.4269	27.6125	29.5870	23.7301	17.8320	26.7327	24.4217	24.6385
21.3333	24.6667	21.2000	27.5714	20.4016	26.2340	19.9703	27.7778
20.3000	32.2857	23.5000	24.7500	17.7953	**40.3571**	25.5385	23.0000
18.1000	24.4444	22.1250	32.3333	17.3028	23.7854	20.8636	33.7736
22.6667	29.0000	20.1667	20.6667	16.1875	31.4012	24.1469	17.7937
18.3000	26.0000	18.2222	23.1250	16.8657	20.8410	25.0917	27.8354
23.0000	30.8333	278.0000	38.3333	15.9153	20.8454	29.8108	20.3881
20.8333	28.0000	25.2000	19.6667	23.0833	30.9688	26.7549	28.8600
22.8000	34.1667	28.0000	31.0000	16.5808	30.7017	25.4951	26.7429
27.6667	22.0000	22.5000	25.8333	16.6300	21.2774	25.7483	26.0123
29.3333	29.6667	23.9000	19.7778	21.6447	25.8104	24.4408	22.4535
17.7368	31.2000	19.1667	29.7500	16.8031	26.7841	20.8571	24.9063
23.4000	23.5000	24.5000	29.6000	20.0772	25.8173	24.7349	23.3158
31.0000	26.7500	19.4286	18.4000	16.2171	24.2424	21.1215	22.5294
19.5000	25.4444	20.5000	23.8571	17.4933	24.2509	25.1731	23.6066
44.5000		21.2857	25.8889	12.3495		22.8227	27.2373
26.2222		29.4444	27.0000	18.4385		23.8649	31.8391
34.0000		20.4286	20.0000	17.1798		24.3537	23.5625
33.2500		24.9091	23.6667	24.0253		24.6220	25.8062
25.8333		24.0000	28.0000	19.8770		30.0833	28.9714
23.2500		19.3750	**50.0000**	17.5000		23.4367	23.8989
16.7500		13.5333	24.0000	24.1585		26.4565	25.0507
25.3750		23.2500	**8.0769**	22.4667		25.1282	26.1474
17.8750		31.5714	16.3333	20.9822		22.0562	16.5263
26.5000		48.2500	28.7143	24.7093		23.8571	24.8188
31.0000			22.5000	19.8704			20.7556
24.1250			32.2000	20.9415			25.8846
20.6000			24.0000	**8.4409**			27.5133
			20.4000				21.2424
			28.5000				26.4156

Table 5.4: Sentence length in the AbstRA corpus (per text)

and one for all RAs. The R function for generating boxplots with notches is `boxplot(mydata, notch=T); grid()`. The resulting plot is displayed in Figure 5.1. R also delivers a summary of the data concerning this plot. Accordingly, abstracts show a minimum sentence length of 8.077 words, 1st quartile of 20.833, median of 24.000, mean of 25.206, 3rd quartile of 28.500, and a maximum sentence length of 50.000 words. Similarly, RAs show the following summary values: minimum sentence length of 8.441, 1st quartile of 20.756, median of 23.865, mean of 23.218, 3rd quartile of 25.810, and a maximum sentence length of 40.357 words. Interestingly, both the minimum and the maximum values for sentence length are found in abstracts from mechanical engineering, which are marked in bold in Table 5.4.

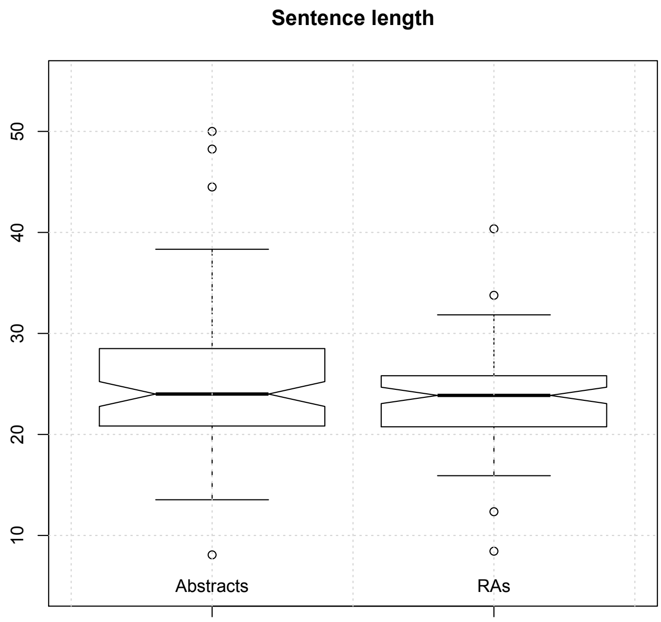


Figure 5.1: Sentence length in the AbstRA corpus

These data comply with data from reference corpora and are in accordance with the expectations for scientific discourse. The mean values for sentence length for the sub-corpus *J* from FLOB and Frown, which contain scientific texts, are 25.47 and 22.54, respectively. The data also indicate that generally abstracts have longer sentences in comparison to their RAs. However, since both notches of the two boxplots apparently overlap (cf. Figure 5.1), there is a possibility that there is *no* significant difference between abstracts and their RAs concerning sentence length. Nevertheless, this assumption has to be tested statistically, as discussed in Section 3.4.2. Before testing for significance, the data has to be tested for normal distribution and homogeneity of variances. Since this is the first feature analyzed according to this methodology, all steps are discussed in detail.

The summary mentioned earlier indicates that the data is *not* normally distributed. For a normally distributed data, the values of the mean and the median are similar. Still, for the sake of completeness, the statistical test for normality, the Shapiro-Wilk test[36], is applied. According to a Shapiro-Wilk test, the distribution of sentence length values in abstracts de-

[36] The R function for the Shapiro-Wilk test is `shapiro.test`.

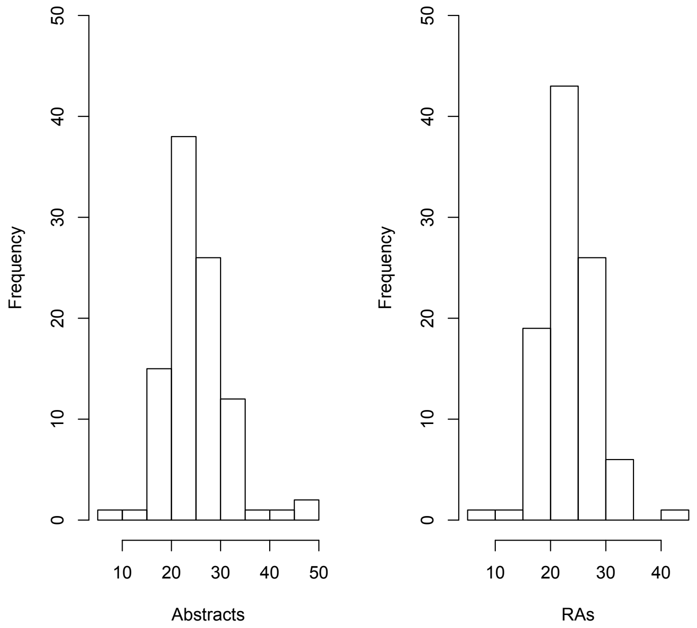


Figure 5.2: Histogram of sentence length values in the AbstRA corpus

viate significantly from normality: W = 0.9162, p-value = 1.186e-05. The same is true for RAs. The distribution of sentence length values in RAs deviate significantly from normality: W = 0.969, p-value = 0.02148. This departure from the normal distribution can also be illustrated in a plot, i.e., histogram[37]. Figure 5.2 shows a histogram for the values of sentence length for both abstracts and RAs in the AbstRA corpus.

The Fligner-Killeen[38] test is the test for homogeneity of variances used here (cf. Section 3.4.2). The Fligner-Killeen test indicates that there is "no compelling evidence for non-constancy of variance" (Crawley 2007: 293) when comparing the variances of the values for sentence length of abstracts and RAs since the returned value in this case is: med chi-squared = 2.8691, df = 1, p-value = 0.0903.

However, because of non-normality, the t-test can not be used and the Wilcoxon rank-sum test[39] must be used (cf. Section 3.4.2). The test is

[37] Histograms are generated by R using the function `hist`.

[38] The R function for the Fligner-Killeen test is `fligner.test`.

[39] The R function for the Wilcoxon rank-sum test is `wilcox.test`, which has an additional parameter called `alternative` allowing for "directional alternative hypotheses"

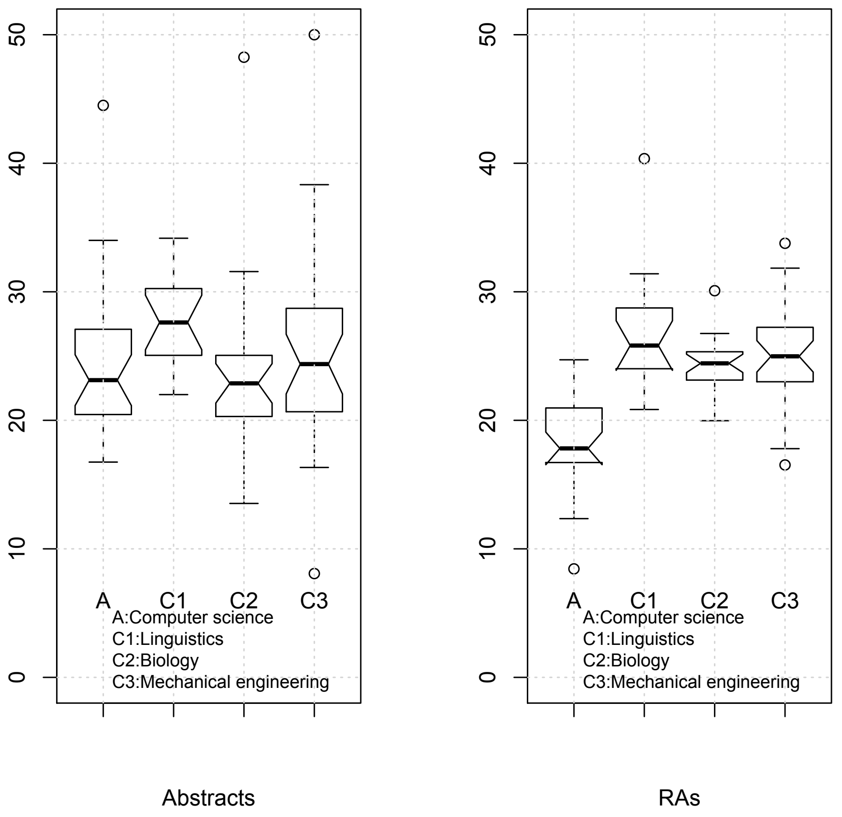


Figure 5.3: Sentence length across disciplines in the AbstRA corpus

performed in the direction that assumes that abstracts do have longer sentences than RAs. The returned value is W = 5451, p-value = 0.02821. Sine the p-value is smaller than 0.05, $\mathbf{H}_0$ is rejected. Hence, abstracts have significantly higher sentence length than their RAs.

Following Gries (2009a: 210), the results for the feature sentence length can be summarized as follows: the median sentence length of abstracts is 24.000 words (interquartile range[40]: 7.67) while the median sentence length of RAs is 23.865 words (interquartile range: 5.06). Since the data violate the assumption of normality, a Wilcoxon rank-sum test is computed. This test shows that the difference between the two sentence lengths is significant (W = 5451, $p_{one-tailed}$ = 0.02821); in the AbstRA corpus, sentences in abstracts are significantly *longer* than in their RAs.

(Gries 2009a: 209). In other words, assuming that abstracts do have longer sentences

The same procedure was followed for the analysis of sentence length across disciplines. Figure 5.3 shows the corresponding boxplot with notches generated from the values in Table 5.4 for abstracts and RAs in each discipline, i.e., computer science (A), linguistics (C1), biology (C2), and mechanical engineering (C3)[41]. According to Figure 5.3, there is variation in sentence length between abstracts and RAs across these four disciplines. Abstracts from linguistics present the longest sentences with a median of 27.61 words, while RAs from computer science show the shortest sentence with a median of only 17.814 words per sentence. When considering only abstracts, Figure 5.3 shows that there is domain specific variation in sentence length since the boxes are vertically differently positioned. For RAs this domain specific variation is also present. However, there are more similarities between biology and mechanical engineering since the medians of the two boxes are very near. In order to investigate the differences between abstracts and their RAs in more detail, each set of abstracts-RAs in a sub-corpus was tested for normality and significance. The corresponding results are presented per discipline.

Computer science

The null hypothesis and the corresponding alternative hypothesis to be tested are:

H$_1$: Abstracts have significantly higher sentence length in comparison to their RAs in the domain of computer science.

H$_0$: Abstracts *do not* have significantly higher sentence length in comparison to their RAs in the domain of computer science.

The median sentence length of abstracts of the discipline of computer science is 23.12 words (interquartile range: 6.27) while the median sentence length of RAs of the discipline of computer science is 17.814 words (interquartile range: 4.19). The data departs from the assumption of normality for abstracts of computer science since the Shapiro-Wilk test

than RAs, this hypothesis can be tested precisely in this direction, using the following code: `wilcox.test(Abstracts, RAs, alternative="greater")`.

[40] Interquartile range measures the statistical dispersion and is equal to the difference between the third and first quartiles.

[41] The abbreviations A for computer science, C1 for linguistics, C2 for biology, and C3 for mechanical engineering is derived from the architecture of the DaSciTex corpus. All "pure" disciplines are denoted with a letter C, all "mixed" disciplines with a letter B (not used in the AbstRA corpus) and computer science, as starting point for discipline comparison, is denoted with a letter A.

indicates W = 0.8964, p-value = 0.009428. Although the data for RAs of computer science do not violate the assumption of normality for the Shapiro-Wilk test returns W = 0.9392, p-value = 0.1055, a Wilcoxon rank-sum test must be computed. This test shows that the difference between the two sentence lengths is significant (W = 641, p-value$_{one-tailed}$ = 2.329e-05). $\mathbf{H}_0$ is thus rejected. For the discipline of *computer science* in the AbstRA corpus, sentences in abstracts are significantly *longer* than in their RAs.

Linguistics

The null and the alternative hypotheses are in this case formulated as follows:

$\mathbf{H}_1$: Abstracts have significantly higher sentence length in comparison to their RAs in the domain of linguistics.

$\mathbf{H}_0$: Abstracts *do not* have significantly higher sentence length in comparison to their RAs in the domain of linguistics.

The median sentence length of abstracts for the discipline of linguistics is 27.61 words (interquartile range: 5.19) while the median sentence length of RAs for the discipline of linguistics is 25.82 words (interquartile range: 4.73). The data conforms to the assumption of normality for abstracts of linguistics since the Shapiro-Wilk test indicates W = 0.9831, p-value = 0.9863. However, the data for RAs in linguistics do not conform to the assumption of normality for the Shapiro-Wilk test returns W = 0.876, p-value = 0.04142. Therefore, a Wilcoxon rank-sum test is computed. This test shows that $\mathbf{H}_0$ *can not* be rejected because the difference between the two sentence lengths is *not* significant (W = 139, p-value$_{one-tailed}$ = 0.1427).

Biology

The alternative and null hypothesis tested are:

$\mathbf{H}_1$: Abstracts have significantly higher sentence length in comparison to their RAs in the domain of biology.

$\mathbf{H}_0$: Abstracts *do not* have significantly higher sentence length in comparison to their RAs in the domain of biology.

The median sentence length of abstracts of the discipline of biology is 22.88 words (interquartile range: 4.62) while the median sentence length of RAs of the discipline of biology is 24.21 words (interquartile range:

1.97). This already indicates a possibility that the null hypothesis is not to be rejected. Nevertheless, the tests have to be performed. The data depart from the assumption of normality for abstracts of biology since the Shapiro-Wilk test indicates W = 0.7996, p-value = 0.0002906. However, the data for RAs of biology conform to the assumption of normality for the Shapiro-Wilk test returns W = 0.9448, p-value = 0.2088. Nevertheless, a Wilcoxon rank-sum test must be computed. This test shows that the difference between the two sentence lengths is *not* significant (W = 215, p-value$_{one-tailed}$ = 0.9347). For this reason, $\mathbf{H}_0$ *can not* be rejected.

Mechanical engineering

The last pair to be compared are abstracts and their RAs in mechanical engineering. The null hypothesis and the corresponding alternative hypothesis to be tested are:

$\mathbf{H}_1$: Abstracts have significantly higher sentence length in comparison to their RAs in the domain of mechanical engineering.

$\mathbf{H}_0$: Abstracts *do not* have significantly higher sentence length in comparison to their RAs in the domain of mechanical engineering.

The median sentence length for abstracts of the discipline of mechanical engineering is 24.375 words (interquartile range: 7.54) while the median sentence length for RAs of the discipline of mechanical engineering is 24.98 words (interquartile range: 4.03). Again, this is already an indication, that H_0 will not probably be refuted. The data depart from the assumption of normality for abstracts of mechanical engineering since the Shapiro-Wilk test indicates W = 0.9181, p-value = 0.02399. However, the data for RAs of mechanical engineering conform to the assumption of normality for the Shapiro-Wilk test returns W = 0.9826, p-value = 0.8895. Nevertheless, a Wilcoxon rank-sum test must be computed. This test shows that the difference between the two sentence lengths is *not* significant (W = 459, p-value$_{one-tailed}$ = 0.45) and $\mathbf{H}_0$ *can not* be rejected.

The analysis of sentence length pairwise across disciplines can thus be summarized as follows: only for the discipline of computer science, abstracts and their RAs differ significantly from each other.

5.1.1.3 Type/token ratio

Type/token ratio (TTR) is defined as “the number of different lexical items in a text, as percentage” (Biber 1988: 238). TTR is a linguistic feature reflecting vocabulary range and, according to Biber, also high density of information.

> [...] [T]ype/token ratio [...] mark[s] high density of information [...] [T]hey further mark very precise lexical choice resulting in an exact presentation of informational content. A high type/token ratio results from the use of many different lexical items in a text, and this more varied vocabulary reflects extensive use of words that have very specific meanings. (Biber 1988: 104)

This feature is computed by WordSmith Tools. Type/token ratio varies very widely in accordance with the length of the text. Therefore, the *standardized* type/token ratio (STTR) is used here. The STTR is computed every *n* words as WordSmith Tools goes through each text file. The minimum possible value of *n* is 100 words and this is the value adopted in this research since abstracts comprise generally less than 200 words. However, there are some abstracts comprising even less than 100 words. For these cases, no STTR is calculated.

It is assumed that abstracts have higher type/token ratio, i.e., *vocabulary range*, in comparison to their RAs, due to their purpose of summarizing information in a small piece of text. Therefore, it is assumed that abstracts do not have many repeated words, otherwise it would decrease the type/token ratio. Instead, abstracts are expected to deploy a great variety of lexical items, i.e., different words, to express the information content of the RA in a summarized form. This assumption is illustrated in Example 5.3, a biology abstract, and in Example 5.4, a RA from computer science:

(5.3) The minichromosome maintenance (MCM) proteins are thought to function as the replicative helicases in eukarya and archaea. The proteins of only a few archaeal organisms have been studied and revealed that although all have similar amino acid sequences and overall structures they differ in their biochemical properties. [abstract.C2.3]

(5.4) It is well known that recognizing of classes of graphs having a k-coloring with a given property is often a hard problem. A proper k-coloring of a graph is a partition of its vertexset into stable subsets. A graph is k-colorable if it has a proper k-coloring. [RA.A.10]

Abstracts				Research articles			
Computer science	Linguistics	Biology	Mechanical engineering	Computer science	Linguistics	Biology	Mechanical engineering
64.1667	67.8000	73.4651	62.4348	60.3885	66.9723	64.7504	62.8288
77.0000	72.0000	66.0000	71.0000	57.8600	68.0000	62.7500	62.5833
68.0000	72.0000	68.0000	56.0000	60.9733	69.8021	64.9375	64.2941
55.0000	65.0000	65.0000	63.0000	55.2985	67.5421	66.6364	64.0588
62.0000	72.0000	63.5000	68.0000	57.8947	**73.4753**	60.7059	61.7273
62.0000	68.0000	77.0000	70.0000	60.3556	63.6250	66.0667	61.8095
67.0000	67.0000	**91.6923**	51.0000	61.7838	66.1618	67.1818	58.5926
59.0000	67.0000	70.0000	64.0000	66.7000	70.9141	69.4074	66.0714
68.0000	66.0000	68.0000	53.0000	62.9630	67.8445	69.8077	60.5000
65.0000	72.0000	69.0000	64.0000	66.2703	64.0161	62.5676	72.8571
62.5000	67.0000	68.5000	61.0000	59.3673	62.9815	62.8649	60.8421
65.3333	62.5000	74.0000	63.0000	59.3023	64.6809	64.1500	62.3478
68.0000		61.0000	61.0000	68.5098	67.2169	66.7500	62.6452
51.0000		68.0000	56.0000	61.1000	56.8723	66.0909	60.5000
64.5000		70.0000	68.0000	56.8462	62.0151	67.0769	64.7143
71.0000		68.0000	55.0000	58.4000		59.7187	65.8125
59.0000		52.0000	66.5000	61.6842		61.8077	61.7778
58.0000		77.0000	77.0000	61.2889		68.3158	59.8667
68.0000		62.5000	63.0000	57.7027		61.0968	61.4848
65.0000		58.5000	65.0000	58.7917		61.0930	59.9000
68.0000		74.0000	58.0000	60.6667		66.7027	63.0952
63.0000		71.0000	55.0000	64.0000		69.0417	59.5588
		57.0000		63.9615		66.3793	64.5000
		60.0000		63.0426		66.0000	57.6667
				63.1429		65.0667	65.6471
				62.1000			70.1667
				60.7692			61.6500
				53.3548			65.9355
							58.1429
							60.3000

Table 5.5: Type/token ratio in the AbstRA corpus (standardized to 100 tokens; per text)

The null hypothesis to be tested, H_0, and its counterpart, the alternative hypothesis H_1, can therefore be formulated as follows:

H_1: Abstracts have significantly higher standardized type/token ratio in comparison to their RAs.

H_0: Abstracts *do not* have significantly higher standardized type/token ratio in comparison to their RAs.

The results of STTR for the AbstRA corpus are summarized in Table 5.5. Abstracts show STTR values varying from 51.00 for computer science up to 91.69 for biology, while RAs have a STTR range between 53.35 for

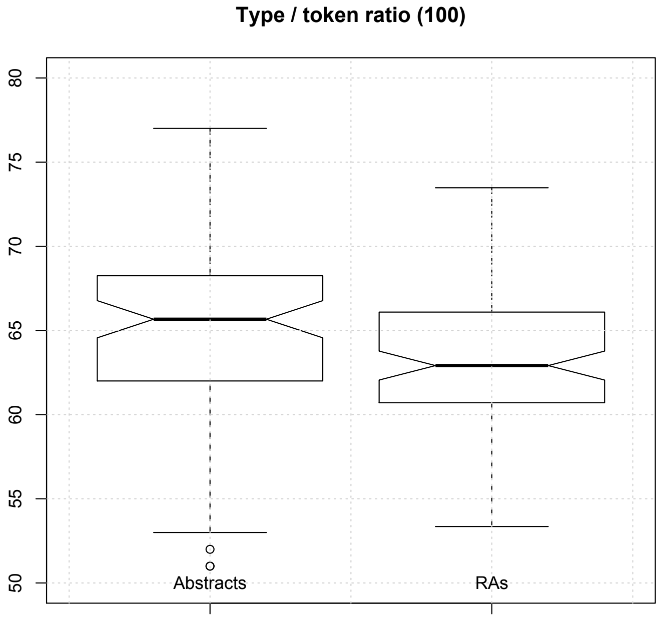


Figure 5.4: Type/token ratio in the AbstRA corpus (standardized to 100 tokens)

computer science and 73.48 for linguistics. The STTR values for AbstRA are mainly higher than the overall STTR values for the scientific texts (*J* subcorpus) of Frown (68.67) und FLOB (68.99). Such results corroborate the expectations for scientific discourse. Moreover, STTR values for AbstRA are higher than the ones found by Steiner et al. (2007: 21) for their corpus of popular-scientific texts (10.98). This also corroborates the initial expectation of higher STTR values indicating wide vocabulary range.

The STTR results for the AbstRA corpus are better visualized through a boxplot with notches, as shown in Figure 5.4. According to this Figure, abstracts show a minimum standardized type/token ratio of 51.00, 1st quartile of 62.00, median of 65.67, mean of 65.41, 3rd quartile of 68.12, and a maximum standardized type/token ratio of 91.69. Similarly, RAs show the following summary values: minimum standardized type/token ratio of 53.35, 1st quartile of 60.72, median of 62.91, mean of 63.32, 3rd quartile of 66.09, and a maximum standardized type/token ratio of 73.48. Again, a first indication that the STTR data are *not* normally distributed is that the values for median and mean are not identical. Therefore, the Shapiro-Wilk test is applied. According to a Shapiro-Wilk test, the distribution

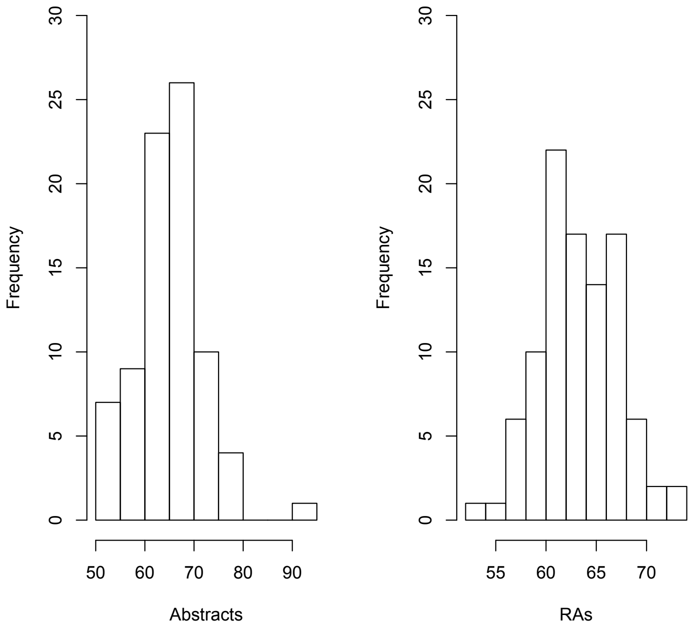


Figure 5.5: Histogram of standardized type/token ratio in the AbstRA corpus

of the values for standardized type/token ratio in abstracts deviate significantly from normality: W = 0.962, p-value = 0.01795. In contrast, the distribution of the values for standardized type/token ratio in RAs does not deviate significantly from normality: W = 0.9929, p-value = 0.8871. These profiles for the distribution of the STTR values are illustrated as histograms in Figure 5.5. The number of STTR values are unfortunately not identical for abstracts and RAs. For some for abstracts, no STTR is calculated because they contain less than 100 words. For this reason, the Fligner-Killeen test for homogeneity of variances can not be applied. However, this causes no further problems in the statistical evaluation of data because the normality pre-requisite for using a t-test is already violated in the abstracts sub-corpus. In other words, the Wilcoxon rank-sum test must be applied anyway for significance testing. Since $\mathbf{H}_0$ is formulated as abstracts having significantly higher STTR than their RAs, the one-tailed Wilcoxon rank-sum test in this direction is applied. The calculated value is W = 4903.5, p-value = 0.002019. Since the p-value is smaller than 0.05, $\mathbf{H}_0$ is rejected. Abstracts have significantly higher standardized type/token ratio than their RAs.

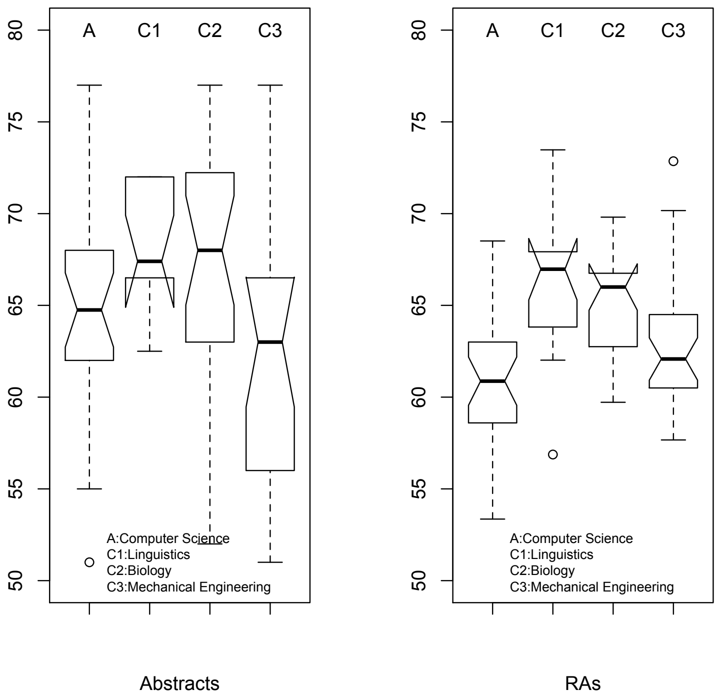


Figure 5.6: Type/token ratio across disciplines in the AbstRA corpus (standardized to 100 tokens)

The results for the feature STTR can be summarized as follows: The median standardized type/token ratio of abstracts is 65.67 (interquartile range: 6.13) while the median standardized type/token ratio of RAs is 62.91 (interquartile range: 5.36). Since the data violate the assumption of normality, a Wilcoxon rank-sum test is computed. This test shows that the difference between the two standardized type/token ratios is significant (W = 4903.5, $p_{one-tailed}$ = 0.002019) indicating that in the AbstRA corpus, standardized type/token ratio in abstracts are significantly *higher* than their RAs.

The same procedure was followed for the analysis of STTR across disciplines. Figure 5.6 shows the corresponding boxplot with notches generated from the values in Table 5.5 for abstracts and RAs in each discipline, i.e., computer science (A), linguistics (C1), biology (C2), and mechanical engineering (C3). According to Figure 5.6, there is a variation in STTR between abstracts and RAs across all four disciplines. Abstracts from biology present the highest STTR with a median of 68.00, while RAs from computer science show the lowest STTR with a median of 60.87. However, the folded notches in Figure 5.6 for abstracts of linguistics and for RAs from linguistics and biology indicate that either the samples are of small size and/or have high within-variance. This warning reinforces the importance of analytical statistic tests for hypothesis testing. When considering only abstracts or only RAs, it can be noticed that there is domain specific variation in STTR since the boxes are vertically differently positioned. Similarly to the procedure followed by sentence length, each set of abstracts-RAs was tested for normality and significance. The results from the Wilcoxon rank-sum test indicate that STTR for abstracts are significantly higher in comparison to their RAs for the disciplines of computer science (W = 445, p-value$_{one-tailed}$ = 0.003801) and biology (W = 392.5, p-value$_{one-tailed}$ = 0.03284) since the p-values > 0.05 and therefore $\mathbf{H}_0$ can be rejected. In contrast, there is no significant difference between the STTR values for abstracts and their RAs for the disciplines of linguistics (W = 115.5, p-value$_{one-tailed}$ = 0.1107) and mechanical engineering (W = 329, p-value$_{one-tailed}$ = 0.5111).

5.1.1.4 Lexical words

The last of the shallow features to be analyzed is the distribution of lexical words in the AbstRA corpus. The linguistic category *lexical word* comprises nouns (and personal pronouns[42]), verbs, adjectives and adverbs. For academic discourse it is generally expected that nouns are the most frequent kind of lexical words indicating *abstractness* (Biber et al. 2002: 23) and *high density of information* (Biber 1988) in texts:

> Nouns are the primary bearers of referential meaning in a text, and a high frequency of nouns thus indicates great density of information.
> (Biber 1988: 104)

Since abstracts summarize the content of RAs, i.e., abstracts tend to incorporate great amount of information in a relative small piece of text, it may be expected that abstracts have a higher frequency of occurrence of nouns than their RAs. Knowing that adjectives modify nouns, it is also expected that the more frequent nouns occur, the more frequent adjectives also occur. Therefore, it can be assumed that adjectives are more frequent in

[42] Personal pronouns refer to “the speaker, the addressee(s), and other entities” (Biber et al. 2002: 26). Personal pronouns are functional words and not lexical words. However, they were included in this section in order to gain first insights into their distribution, since they stand for nouns in text and no other functional words are to be investigated in this study.

abstracts than RAs. Example 5.5, from an abstract from computer science, and Example 5.6, a linguistic RA elucidate this expectation:

(5.5) In this paper, a new method for handling multicriteria fuzzy decision-making problems based on intuitionistic fuzzy sets is presented. The proposed method allows the degrees of satisfiability and non-satisfiability of each alternative with respect to a set of criteria to be represented by intuitionistic fuzzy sets, respectively. [abstract.A.15]

(5.6) In these remarks, I will try to identify what seem to me some of the significant themes in the past half-century of inquiry into problems of biolinguistics and to consider their current status. Several preliminary qualifications should be obvious. [abstract.C1.3]

Therefore, the hypotheses to be tested in this section are:

H$_1$: Abstracts show a significantly higher frequency of occurrence of nouns and adjectives in comparison to their RAs.

H$_0$: Abstracts *do not* show a significantly higher frequency of occurrence of nouns and adjectives in comparison to their RAs.

Table A.1 (Appendix A.3, p. 202) and Table A.2 (Appendix A.3, p. 204) show the distribution of lexical words for abstracts and RAs in each discipline, respectively. According to Tables A.1 and A.2, nouns are undoubtedly the most common types of lexical words both in abstracts and in RAs. The frequency of occurrence of nouns in abstracts varies from a minimum of 18.68% to a maximum of 40% of all tokens, with a median of 30.65%, while in RAs this range is from 20.63% to 35.56%, with a median of 27.99%.

For better data visualization and interpretation, Figure 5.7 shows the same data for distribution of lexical words in the AbstRA corpus as a boxplot with notches. According to Figure 5.7, nouns seems to be the lexical word with the most significant difference in the frequency of occurrence between abstracts and RAs.

The Shapiro-Wilk test is applied to test the normal distribution of nouns in abstracts and RAs. Accordingly, the test shows that the distribution of the values for frequency of occurrence of nouns in abstracts do not deviate from the normal distribution since W = 0.9919, p-value = 0.846. This is also true for RAs, where the Saphiro-Wilk test returns the value W = 0.9843, p-value = 0.3225. Since the data do not conform to the prerequisite of normality, a *t-test* for independent samples can be applied to test for significance. The t-test is performed in R with the function `t.test`.

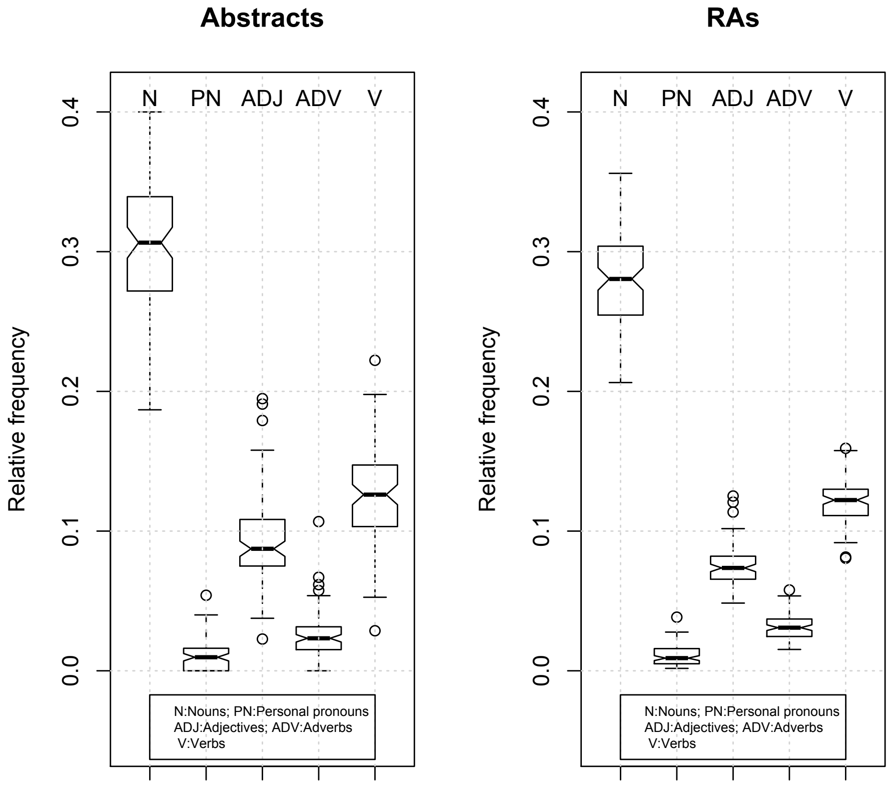


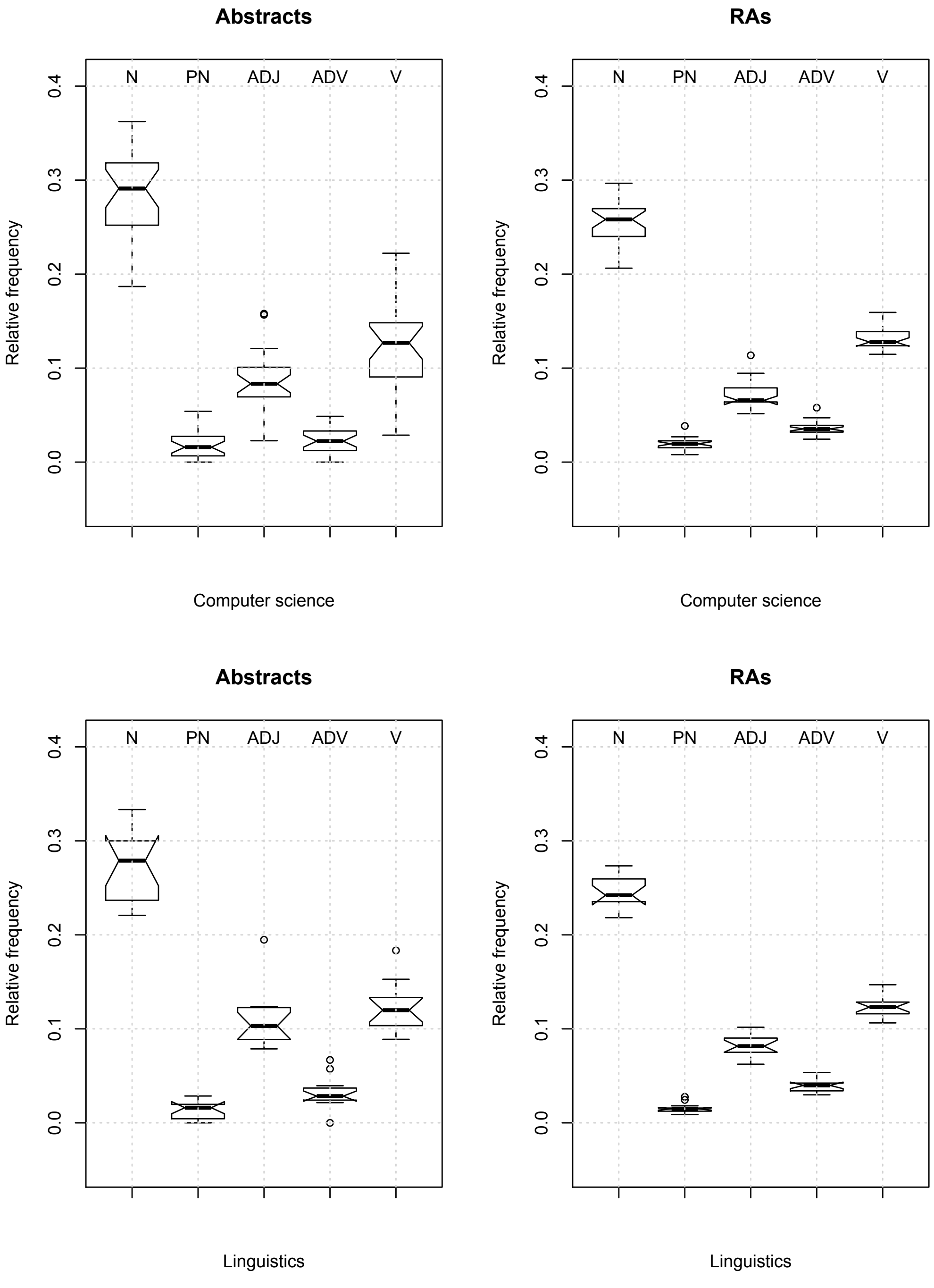
Figure 5.7: Distribution of lexical words in the AbstRA corpus (relative frequencies)

It shows that the difference between the two frequencies of occurrences of nouns is very significant (t = 4.2962, df = 174.063, p-value$_{one-tailed}$ = 1.442e-05): the first part of $\mathbf{H}_0$ can be rejected. Thus, abstracts show significantly higher frequency of occurrence of nouns in comparison to their RAs in the AbstRA corpus. The same tests are performed for adjectives, personal pronouns, adverbs and verbs. The data do not conform to the pre-requisite of normality in all these cases. Therefore, only Wilcoxon rank-sum test are performed for significance testing. Adjectives are significantly more frequent in abstracts than in their RAs (W = 6129, p-value$_{one-tailed}$ = 1.022e-06). Therefore, the second part of $\mathbf{H}_0$ can also be rejected. However, there is no significant difference in the frequencies of occurrence of personal pronouns in abstracts and their RAs (W = 4035, p-value$_{one-tailed}$ = 0.8187). In contrast, adverbs are significantly more frequent in RAs than in abstracts (W = 2780, p-value$_{one-tailed}$ = 8.622e-06; W = 2780, p-value = 1.724e-05).

Finally, there is no significant difference in the frequencies of occurrence of verbs in abstracts in comparison to their RAs (W = 4709, p-value$_{one-tailed}$ = 0.1809).

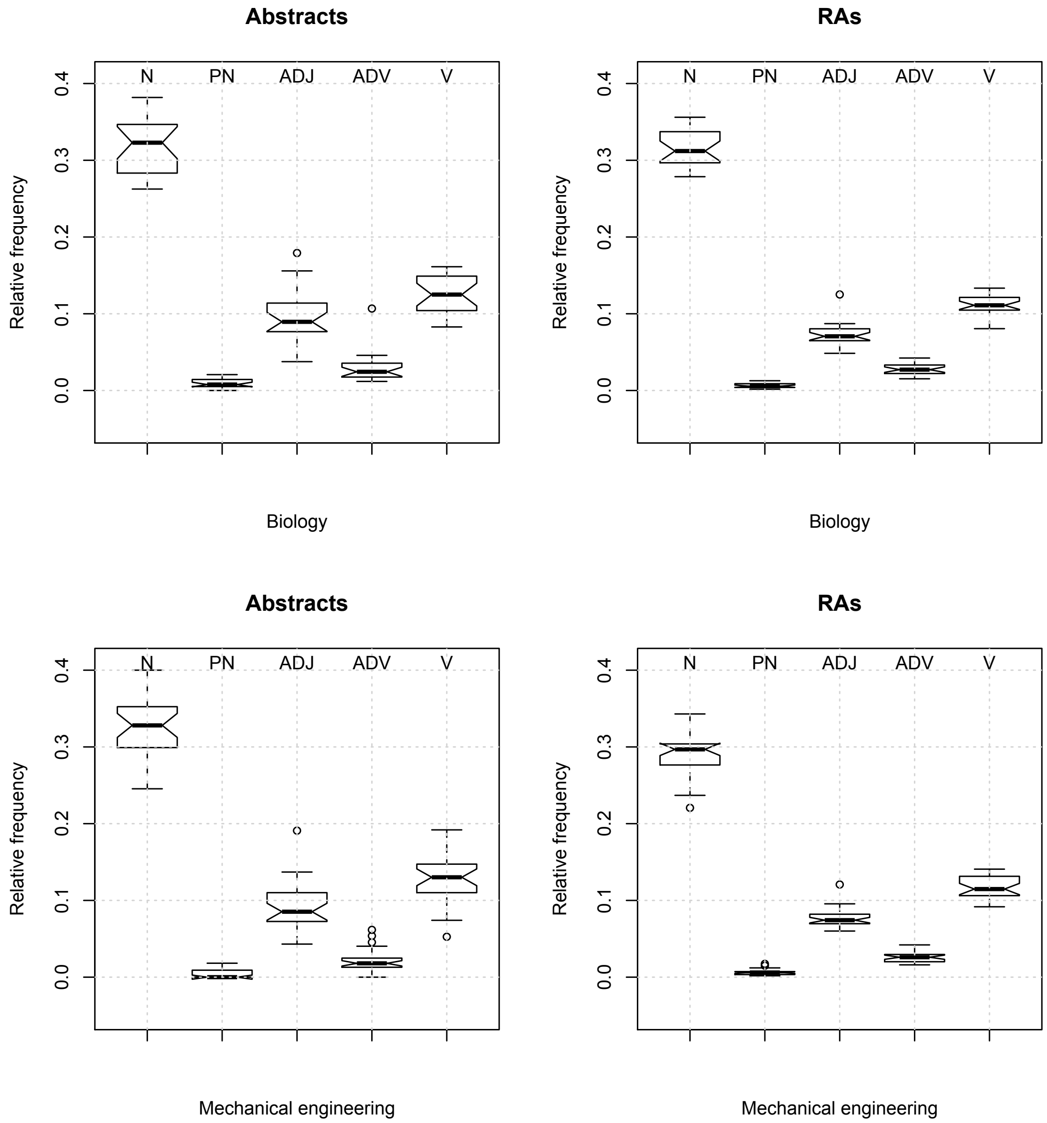
The next step is to compare the distribution of lexical words pairwise across domains. Figures 5.8 and 5.9 show the corresponding boxplots for the abstracts-RAs pairs in the disciplines of computer science, linguistics, biology, and mechanical engineering. When considering only abstracts or only RAs, Figures 5.8 and 5.9 show that there is domain specific variation in the distribution of lexical words since the boxes are vertically differently positioned, specially for nouns and verbs. However, the tests for significance are only performed for nouns and adjectives since they are the lexical words that showed significant difference between abstracts and RAs in the AbstRA corpus, also across disciplines. Again, data is not normally distributed. The Wilcoxon rank-sum test shows that adjective and nouns in *computer science* are significantly more frequent in abstracts than in their RAs (W$_{Nouns}$ = 531, p-value$_{one-tailed}$ = 0.002040; W$_{Adjectives}$ = 494, p-value$_{one-tailed}$ = 0.01239). The discipline of *linguistics* shows a similar profile, where abstracts use both nouns and adjectives more frequently than their RAs (W$_{Nouns}$ = 141, p-value$_{one-tailed}$ = 0.02487; W$_{Adjectives}$ = 166, p-value$_{one-tailed}$ = 0.0005914. In contrast, there is no significant difference in the frequencies of occurrence of nouns in *biology* since W = 301, p-value$_{one-tailed}$ = 0.3027. Nevertheless, adjectives are significantly more frequent in abstracts than in their RAs (W = 407, p-value$_{one-tailed}$ = 0.002372). Finally, the Wilcoxon rank-sum test shows that adjective and nouns in *mechanical engineering* are significantly more frequent in abstracts than in their RAs (W$_{Nouns}$ = 640, p-value$_{one-tailed}$ = 0.0003298; W$_{Adjectives}$ = 540, p-value$_{one-tailed}$ = 0.03211).

The analysis of the shallow features allow the researcher to gain first insight into the linguistic properties of the AbstRA corpus. All shallow features analyzed show overall significant differences between abstracts and RAs as well as very often across individual domains. The results corroborate so far the working hypotheses that abstracts are significantly different in comparison to their RAs and support further the choice of linguistic features for the quantitative analysis. Section 5.1.2 discusses the results for the chosen lexical features.



N:Nouns; PN: Personal pronouns; ADJ: Adjectives; ADV; Adverbs, V:Verbs

Figure 5.8: Distribution of lexical words for computer science and linguistics in the AbstRA corpus (relative frequencies)



N:Nouns; PN: Personal pronouns; ADJ: Adjectives; ADV; Adverbs, V:Verbs

Figure 5.9: Distribution of lexical words for biology and mechanical engineering in the AbstRA corpus (relative frequencies)

5.1.2 Lexical features

This section aims to discuss the results concerning the three lexical features chosen for quantitative analysis of the AbstRA corpus. First, the distribution of lexical density is addressed in Section 5.1.2.1. Then, the most frequent lexical items are presented and discussed in Section 5.1.2.2, followed by an analysis of their keyness in comparison to the BNC in Section 5.1.2.3. It must be noted that due to the size of the AbstRA corpus, the results of Sections 5.1.2.2 and Section 5.1.2.3 only represent the data present in the corpus under study. They should not be used as an extrapolation for representativeness in scientific discourse as a whole since such studies demand corpora with at least one million words. However, the obtained data reflects a given tendency in abstracts and RAs.

5.1.2.1 Lexical density

Lexical density measures the density of information in a text, "according to how tightly the lexical items have been packed into the grammatical structure" (Halliday 1993b: 76). There are several methods for measuring lexical density (cf. Halliday 1993b; Stubbs 1986; Ure 1971). The method used in this research is the one suggested by Halliday, who defines lexical density as "the number of lexical words per clause" (Halliday 1993b: 76). According to Halliday, texts even become difficult to read if the values for lexical density are higher than 10, i.e., with more than 10 lexical words per clause.

Although there is no exact information about the clause boundaries and the number of clauses in the AbstRA corpus, lexical density can be calculated assuming that *each* main clause has *one* finite verb. This is a valid approximation which only considers finite clauses for the calculation of lexical density. Non-finite clauses are therefore not taken into consideration in the calculation of lexical density here.

Lexical density can thus be calculated based on the distribution of the PoS-tags for verbs concerning finiteness. The list of tags with corresponding part-of-speech can be found in Appendix A.2 (p. 199). In order to decide whether a verb-tag is finite or not, queries in IMS-CWB/CQP are performed (cf. Section 4.2). According to the list on Appendix A.2 and the results of the queries in CQP, the finite verb tags are VBD, VBP, VBZ, VHD, VHP, VHZ, VVD, VVP and VVZ, while the non-finite verb tags are VB, VBG, VBN, VH, VHG, VHN, VV, VVG, and VVN. Similarly, the lexical words needed for the calculation of lexical density, i.e., cardinals, nouns, adjectives, adverbs, can also be taken from the distribution of the PoS-tags since they are tagged as CD*, N*, J*, and R*, respectively. Finally, the lexical verbs are equal to the total number of verbs minus the total amount of all verbs minus auxiliaries. Thus, the formula for calculating the value for lexical density (LD) for each single text in the AbstRA corpus is:

$$LD = \frac{\sum \textit{lexical words}}{\sum \textit{clauses}}$$

$$LD = \frac{\sum (cardinals + nouns + adjectives + adverbs + lexical\ verbs)}{\sum finite\ verbs}$$

$$LD = \frac{\sum (CD + N^* + J^* + R^* + (V^* - (VBG + VBN + VHG + VHN)))}{\sum (V^* - (VB + VBG + VBN + VH + VHG + VHN + VV + VVG + VVN))}$$

where

All verbs = V* = VB + VBD + VBG + VBN + VBP + VBZ + VH + VHD + VHG + VHN + VHP + VHZ + VV + VVD + VVG + VVN + VVP + VVZ;

Lexical verbs = (V* - (VBG + VBN + VHG + VHN)) = all verbs minus auxiliaries (be, have in passive voice or functioning as participle) and modals (they are tagged separately as MD);

Finite Verbs = (V* - (VB + VBG + VBN + VH + VHG + VHN + VV + VVG + VVN) = all verbs minus total-non-finite verbs.

Taking into consideration that the main purpose of abstracts is to summarize the knowledge of the whole RA, it is reasonable to assume that the lexical density in abstracts are higher than in their RAs. Example 5.7, from an abstract from mechanical engineering with a very high lexical density, and Example 5.8, a linguistic RA with a much lower lexical density illustrate this assumption:

(5.7) The thin coating on the sand bed particles from full-scale boilers was found to have a significant effect on the heat transfer coefficient, while the particle size distributions, as well as coating thickness, had little or no influence on the heat transfer coefficients for the conditions investigated. [abstract.C3.3]

(5.8) In my work on Theme I have found it useful to deal with a unit slightly larger than clause, but smaller than sentence. [RA.C1.1]

Abstracts				Research articles			
Computer science	Linguistics	Biology	Mechanical engineering	Computer science	Linguistics	Biology	Mechanical engineering
18.0	17.8	18.1	11.2	9.1	9.1	13.2	11.4
6.9	9.9	12.0	12.3	8.7	11.5	11.2	13.5
24.0	11.8	7.7	13.8	8.6	9.7	11.6	12.4
10.2	10.5	9.6	13.5	8.1	8.2	14.7	11.4
7.8	6.6	15.5	8.8	8.8	8.2	13.6	12.1
9.9	8.4	9.9	14.8	11.8	8.7	**17.6**	13.2
16.0	7.5	8.9	8.9	8.2	8.2	11.2	9.6
10.9	10.8	9.4	7.8	9.6	8.1	11.2	12.4
11.5	8.6	14.3	15.2	8.6	8.9	11.1	12.0
8.6	12.1	10.8	14.1	7.9	7.8	13.6	12.5
12.3	7.7	14.6	12.6	8.4	10.0	12.6	9.7
10.8	9.7	8.9	11.1	8.0	8.2	13.8	12.9
11.0	12.6	12.3	12.2	8.8	9.9	10.4	11.0
5.6	9.2	12.4	9.7	7.2	9.8	12.2	12.3
19.7		7.1	7.4	10.2	7.5	10.4	8.7
6.7		15.9	9.3	7.3		11.9	13.7
10.1		14.5	15.0	11.8		13.1	14.8
7.8		8.8	13.7	7.6		13.4	16.4
24.8		7.5	7.8	9.0		11.5	8.7
8.9		13.0	**26.0**	10.0		11.5	13.3
13.3		11.0	13.9	10.6		11.0	10.6
15.4		13.4	13.6	9.8		14.1	13.1
10.5		10.3	15.3	10.4		11.4	7.9
16.0			11.8	**7.1**		12.0	12.0
10.3			11.0	7.6			13.2
12.7			10.4	7.4			10.7
11.1			10.1	10.7			11.4
			12.6				12.1
			11.3				9.7

Table 5.6: Lexical density in the AbstRA corpus (per text)

Thus, the null hypothesis to be tested, H_0, and its counterpart, the alternative hypothesis H_1, can be formulated as follows:

H_1: Abstracts have significantly higher lexical density in comparison to their RAs.

H_0: Abstracts *do not* have significantly higher lexical density in comparison to their RAs.

Table 5.6 presents the corresponding results for lexical density in the AbstRA corpus. The values of lexical density for abstracts vary from a minimum of 5.60, for abstracts of computer science, to a maximum of 26.00, for abstracts of mechanical engineering. RAs show values for lexical density from a minimum of 7.10, for a RA of computer science, to a maximum

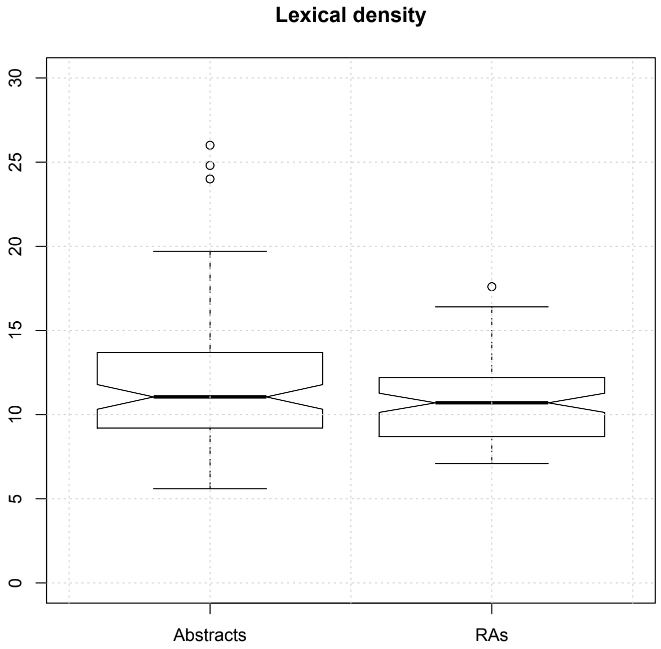


Figure 5.10: Lexical density in the AbstRA corpus

of 17.60, for a RA of biology. The values for lexical density in the AbstRA corpus are mostly higher than the overall lexical density values for the scientific texts (*J* subcorpus) of Frown (9.82) und FLOB (9.39). Such results corroborate the expectations for scientific discourse and they show a tendency for abstracts having a higher lexical density than RAs.

The results for lexical density in the AbstRA corpus are better visualized through a boxplot with notches, as shown in Figure 5.10. According to this Figure, abstracts show a minimum lexical density of 7.10, 1st quartile of 8.73, median of 10.70, mean of 10.73, 3rd quartile of 12.18, and a maximum lexical density of 17.60. Similarly, RAs show the following summary values: minimum lexical density of 5.60, 1st quartile of 9.20, median of 11.00, mean of 11.78, 3rd quartile of 13.60, and a maximum lexical density of 26.00.

The next step is the statistical evaluation of the data. An indication that the lexical density data are *not* normally distributed is that the values for median and mean are not identical. The Shapiro-Wilk test for normality testing is thus applied. According to a Shapiro-Wilk test, the distribution of the values for lexical density in abstracts deviate significantly from normality: W = 0.9, p-value = 2.952e-06. The same is valid for RAs. Their lexical

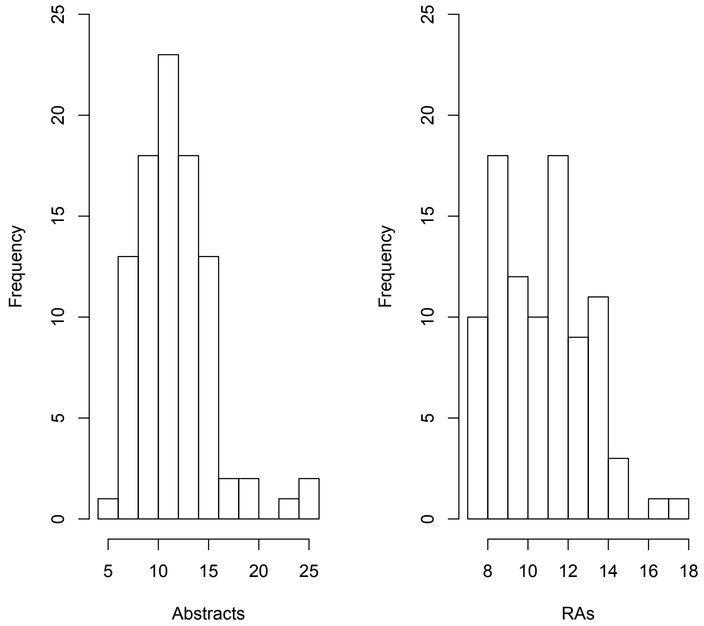


Figure 5.11: Histogram of lexical density in the AbstRA corpus

density values are not normally distributed since W = 0.9672, p-value = 0.01842. These non-normality of data is presented as histograms in Figure 5.11.

Since the pre-requisite for normality for using a t-test was not met in the abstracts sub-corpus, a Wilcoxon rank-sum test must be applied to test for significance. Because the $\mathbf{H}_1$ formulated states that abstracts have significantly higher lexical density than their RAs, the one-tailed Wilcoxon rank-sum test in this direction is applied. The calculated value is W = 4946.5, p-value = 0.0601. This means that there is a 93.99% probability that this difference is *not* due to chance. Furthermore, $\mathbf{H}_0$ *can not* be rejected because the p-value is higher than 0.05, although being very near to it. It should be borne in mind however, that this value is near the border of acceptance, showing a tendency for abstracts having higher values for lexical density than RAs.

Again, the same procedure was followed for the analysis of lexical density across disciplines. Figure 5.12 shows the boxplot with notches generated from the values in Table 5.6 for abstracts and RAs in each discipline, i.e., computer science (A), linguistics (C1), biology (C2), and mechanical

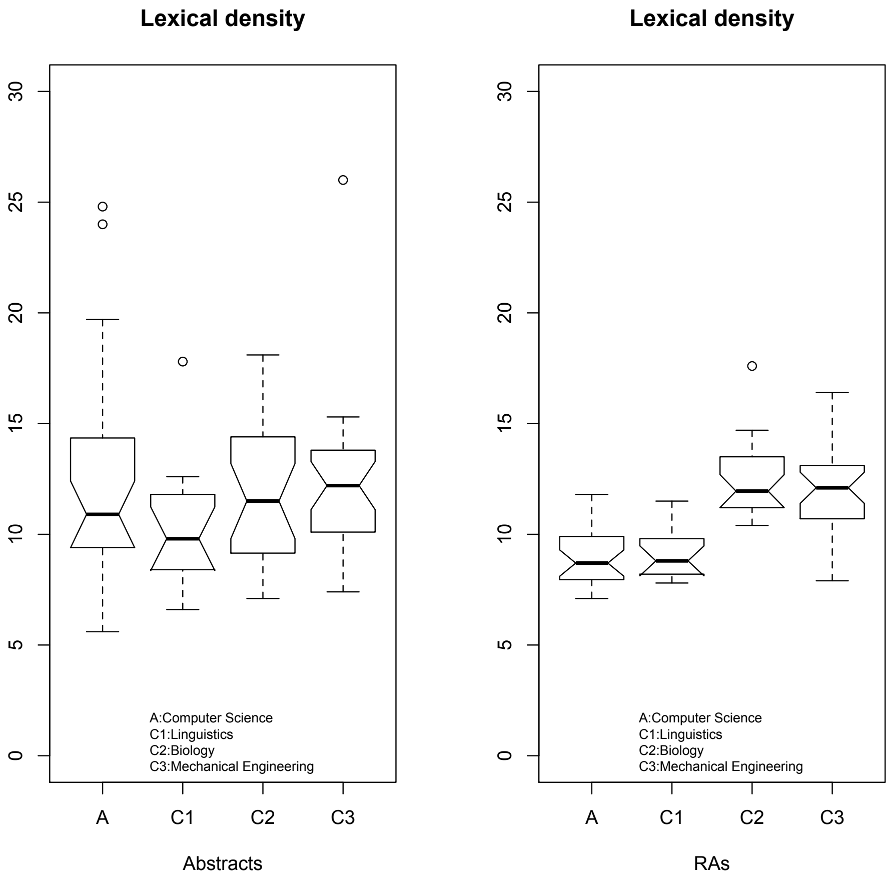


Figure 5.12: Lexical density across disciplines in the AbstRA corpus

engineering (C3). According to Figure 5.12, there is variation in lexical density between abstracts and RAs across all four disciplines. Abstracts from mechanical engineering present the highest lexical density with a median of 12.20, while RAs from computer science show the lowest values with a median of 8.70 lexical words per clause. When considering only abstracts, it can be noticed that there is domain specific variation in lexical density since the boxes are vertically differently positioned. For RAs this domain specific variation is still present, but there are more similarities between computer science and linguistic, as well as between biology and mechanical engineering. As for all shallow features, each pair of abstracts-RAs was tested for normality and significance. Again, most of the data is not normally distributed, which requires the use of a Wilcoxon rank-sum test for significance. The results from the Wilcoxon rank-sum test indicate that lexical density for abstracts is significantly higher in comparison to their RAs only for the discipline of computer science (W = 542.5, p-value$_{one-tailed}$ = 0.001066). For computer science, $\mathbf{H}_0$ can be rejected. In contrast, there is no significant difference between the lexical density values for abstracts and RAs for the disciplines of linguistics (W = 124, p-value$_{one-tailed}$ = 0.1203), biology (W = 206.5, p-value$_{one-tailed}$ = 0.8571) and mechanical engineering (W = 445.5, p-value$_{one-tailed}$ = 0.3516). The significance level is established at 0.05, i.e., 5 percent or greater, for allowing the rejection of the null hypothesis, which in this case is not true. Therefore, although it can *not* be said that abstracts have a *significant* higher lexical density in comparison to RAs across these three disciplines, there is *still* variation in the profile of this feature across disciplines when considering only abstracts or only RAs. This observation is valid for all cases and features where the null hypothesis could not be rejected.

5.1.2.2 Most frequent lexical items

Variation in vocabulary is a well-stablished parameter for investigating variation in language since lexical items are a reflection of the lexical domain and ultimately the field of discourse they represent (cf. Sections 2.4.2 and 4.3). Therefore, in accordance with the working hypothesis of this research that abstracts and RAs differ from each other, it can be expected that the most frequent lexical items occurring in abstracts differ from the ones in RAs, as well as across disciplines.

The identification and quantification of the most frequent lexical items is based on PoS-tags and is performed using WordSmith Tools. The most frequent lexical items, i.e., nouns, adjectives, adverbs, and verbs for abstracts and RAs in the AbstRA corpus are presented in Tables 5.7 and 5.8, respectively. These tables show the twenty most frequent lexical items, their raw frequency of occurrence followed by the relative frequency in percentage. Due to the size of the AbstRA corpus, which is far less from the required minimum of 1 million words for lexical analysis, the findings reported in this section intend to give just some insights in lexical variation in the corpus . Consequently, this feature is not measured by each single text individually and the data are not statistically evaluated unlike the previous features. Nevertheless, taking a look at lexical profiles may reveal relevant information about the corpus under study.

The most frequent *noun* in all abstracts is *problem*, followed by *analysis*, *space*, *time*, *model*, *results*, and *process*. This indicates that the main issue addressed in abstracts is probably a given problem, the model and analysis for investigating such a problem, probably in a reasonable amount of time (and space) and finally the results obtained (cf. Table 5.7). As shown in Table 5.8, research articles show a different profile of the most frequent nouns, i.e., *time*, *number*, *packets*, *analysis*, *case*, *problem*. Therefore, *analysis* and *problem* play a less important role in RAs than in abstracts. Example 5.9, from an abstract from computer science and Example 5.10, a biology abstract illustrate the context of occurrence of some of the most frequent nouns:

(5.9) In addition, [. . .], we obtain a constant O(1/[delta])-approximation ratio for the *problem*. Our *results* have implications for network design. [abstract.A.11; emphasis added]

(5.10) Transcriptome *analysis* can provide useful data for refining genome sequence annotation. [abstract.C2.10; emphasis added]

While in abstracts the most frequent nouns have more of a general meaning, RAs present some domain specific terminology under the top 10, like *number*, *packets* and *algorithm*. Moreover, the addressed problem itself seems to loose importance in RAs since the relative frequency of occurrence of *problem* in RAs (0.1218%) is half of the number of its occurrence in abstracts (0.2452%).

For *adjectives*, which modify nouns, it can be observed that they have an important function in abstracts. Adjectives clarify the uniqueness of a given research. They emphasize what is different and new in a given research in comparison to others since the most frequent adjectives in abstracts are *different* (0.1291%) and *other* (0.1097%) and *new* (0.1033%). The contrastive aspects of a given research are also very important in RAs since the frequency of occurrence of *other* (0.1667%) and *different* (0.1204%) are similar to abstracts. However, the novelty aspect of a research seems to lose importance in RAs, as the frequency of occurrence of *new* (0.0629%) is much lower in RAs than in abstracts. This observation is illustrated in Example 5.11, from an abstract from biology and Example 5.12, a RA from linguistics:

(5.11) We describe here a *new* solution structure of the RNA dimerization initiation site (DIS) of HIV-1Lai. [abstract.C2.20; emphasis added]

(5.12) If there is a contrast in acceptability, such theories must be enriched by further assumptions, such as connectedness theory (Kayne 1983) and *other* approaches looking at movement paths (Pesetsky 1982), or the Minimal Compliance Principle (Richards 2001) claiming that the Superiority Condition need not be respected by more than one pair of wh-phrases in each clause (see also Pesetsky 2000). [RA.C1.11; emphasis added]

To address aspects of similarity and contrast seems to be the main intended purposes for the use of *adverbs* in abstracts and in RAs. The most frequent adverbs in abstracts are *also* (0.1872%), *not* (0.1162%), and *as* (0.1097%). Research articles show a similar profile in the use of adverbs. However, contrastive aspects seem to play an important role since the most frequent adverb in RAs is *not* (0.3565%), occurring almost three times more often than in abstracts. This observation is illustrated in Examples 5.13 and 5.14:

(5.13) We give a matching linear lower bound on the maximum delay incurred by the packets. We *also give* an almost matching linear lower bound on the maximum buffer size used by LIS on DAGs. [RA.A.4; emphasis added]

(5.14) Since the main aim is to find possible analytical solutions but *not* solutions for given initial and boundary conditions, the derivation approach in this paper is different from the common method. [RA.C3.22; emphasis added]

Finally, undoubtedly the most frequent *verbs* in abstracts and RAs are *be* and *have*, as expected[43]. Nevertheless, they behave very differently when it comes to the most frequent verbs other than these. According to SFL, verbs can be classified into six categories according to the kind of processes they realize linguistically. There are material processes which describe actual physical actions. Mental processes describe mental experience. Relational processes are processes of identification and classification. Behavioral processes represent "outer manifestations of inner workings, the acting out of processes of consciousness" (Halliday 2004a: 171). Verbal processes are processes of saying. Finally, existential processes deal with existence in which phenomena are recognized "to be". The most frequent verbs in abstracts, besides *be* and *have*, are therefore either of the relational (*base, show, give,*

[43] Wordlists were retrieved based on the part-of-speech tags. Therefore, different functional roles of the instances of the verbs *be* and *have* (full verb or auxiliary) were not investigated.

contain, present), material (*use, find, obtain, make*), verbal (*propose*) or mental (*investigate*) type. Research articles show some differences in comparison to abstracts, e.g., the absence of verbs of the verbal type under the most frequent ones. The most frequent verbs in RAs besides *be* and *have* are either material (*use, obtain, find*), relational (*give, show, let, follow*), and mental (*see*). Examples 5.15 to 5.19 illustrate the use of different process types in the corpus:

(5.15) In this paper, we *present* a more effective method of computation based upon a 4-state two-dimensional ACA [...] [abstract.A.23; emphasis added]

(5.16) In order to *obtain* high yield of MWCNTs, 900 [deg]C was appropriate instead. [abstract.C3.7; emphasis added]

(5.17) In this paper we *propose* a novel concept to use in the modeling of real network scenarios under measurement and analysis. [abstract.A.6; emphasis added]

(5.18) Finally, we *investigate* the approximability of several extensions of the load rebalancing model. [abstract.A.20; emphasis added]

(5.19) Based on the test results, one can *see* considerable influence of inclination angle and the number of tube row. [RA.C3.5; emphasis added]

A similar evaluation is performed with the ten most frequent lexical items for abstracts and RAs for each single discipline, which are shown in Tables 5.9, 5.10, 5.11, and 5.12. *Problem* is again the most frequent noun in abstracts from computer science. Interestingly, almost all other very frequent nouns in computer science are domain specific terms, e.g., *algorithm*, *graph.* The frequency of the word K (0.5202%), which denote variables in the abstracts, is an indication of a highly formalized and abstract domain. This complies with the observation concerning verbs in abstracts of computer science; they are mainly relational (cf. Table 5.9). Linguistic abstracts contain under the most frequent nouns typical domain specific terms, e.g., *theme, language, unit, sentence.* Interestingly, *argue*, which is a verb of the verbal type, is the most frequent verb besides "be" in linguistics (cf. Table 5.9). Abstracts from biology show only domain specific terms under the top 10 nouns, e.g., *gene*, *DNA*, *methylation.* The focus of the abstract topics on genetics is a consequence of the architecture of the AbstRA corpus, as discussed in Section 4.1. Adjectives in abstracts of biology indicate strong use of noun compounds composed of an adjective and a noun since some examples of very frequent adjectives are *mitochondrial, genomic, human, binding.* Furthermore, verbs in biology are mainly relational or material (cf. Table 5.10). Finally, the most frequent nouns in abstracts from mechanical engineering are also very typical of this domain, i.e., *process, heat, concentration, rate, transfer, flow.* Specially interesting are the most frequent adjectives in abstracts of mechanical engineering; four of the most frequent ones end in the suffix “al”, indicating frequent use of adjective-nouns compounds as domain specific terms in this discipline.

Research articles also show similar variation of the most frequent lexical items across these four disciplines. It is worth mentioning that the most frequent verb besides “be” in computer science is *let*, which indicates a high degree of formality in discourse in this discipline. In contrast, RAs from linguistics use mental verbs very frequently, e.g., *see* (cf. Table 5.11). Mechanical engineering employ in its RAs mainly material verbs besides “be” (cf. Table 5.12).

Research articles from biology show a striking phenomena. This can be observed on the frequent use of *et* and *al*, originally “et al.”. As a matter of fact, the high frequencies of occurrence for these two words are the result of a tagging error. Researchers should be aware that automatic taggers are not free of errors (cf. Section 4.2) and in this case it would be more appropriate for the tagger to treat *et* and *al* as a multi-word unit “et al.”. However the tagger applied here does not consider multi-word units causing problems in the tagging results of scientific discourse. Strikingly, the frequency of occurrence of these two words are however not identical. This is because there is a cited author whose surname is just *al.* This information is shown in Table 5.12 only as a warning for researchers. These data are not taken into consideration in the interpretation of the results.

The analysis of the most frequent items has so far allowed the researcher to gain insights into the semantic content of abstracts and RAs as well as its domain specific variation. However, the data reveal no information so far whether the occurrence of these most frequent lexical items is higher than expected for a corpus of general English or not. One method for uncovering such information is called *analysis of keywords*, which is discussed in the next section.

Nouns			Adjectives			Adverbs			Verbs		
Word	Freq	%	Word	Freq	%	Word	Freq	%	Word	Freq	%
PROBLEM	38	0.2452	DIFFERENT	20	0.1291	ALSO	29	0.1872	IS	211	1.3617
ANALYSIS	30	0.1936	OTHER	17	0.1097	NOT	18	0.1162	ARE	99	0.6389
SPACE	30	0.1936	NEW	16	0.1033	AS	17	0.1097	WAS	61	0.3937
TIME	30	0.1936	MITOCHONDRIAL	15	0.0968	HERE	16	0.1033	BE	48	0.3098
MODEL	27	0.1742	CONSTANT	14	0.0904	HOWEVER	16	0.1033	WERE	40	0.2581
RESULTS	27	0.1742	HIGH	14	0.0904	ONLY	15	0.0968	HAVE	35	0.2259
PROCESS	26	0.1678	LINEAR	14	0.0904	VERY	10	0.0645	BEEN	34	0.2194
GENE	24	0.1549	FUZZY	12	0.0774	WELL	8	0.0516	HAS	27	0.1742
PROTEINS	24	0.1549	MULTIPLE	12	0.0774	MORE	7	0.0452	BASED	25	0.1613
TEMPERATURE	23	0.1484	EXPERIMENTAL	11	0.0710	RESPECTIVELY	7	0.0452	USING	23	0.1484
HEAT	20	0.1291	HUMAN	10	0.0645	THEN	7	0.0452	SHOW	20	0.1291
PROTEIN	20	0.1291	SUCH	10	0.0645	FURTHERMORE	6	0.0387	GIVEN	18	0.1162
METHOD	19	0.1226	INTUITIONISTIC	9	0.0581	EVEN	5	0.0323	USED	18	0.1162
PAPER	19	0.1226	USEFUL	9	0.0581	FINALLY	5	0.0323	FOUND	16	0.1033
ALGORITHM	18	0.1162	FUNCTIONAL	8	0.0516	IN	5	0.0323	PROPOSED	14	0.0904
CELLS	18	0.1162	GENOMIC	8	0.0516	OFTEN	5	0.0323	CONTAINING	12	0.0774
DNA	18	0.1162	HIGHER	8	0.0516	POTENTIALLY	5	0.0323	OBTAINED	12	0.0774
NETWORK	18	0.1162	IMPORTANT	8	0.0516	SO	5	0.0323	INVESTIGATED	10	0.0645
SEQUENCE	18	0.1162	PHI	8	0.0516	BEST	4	0.0258	MADE	9	0.0581
RATE	17	0.1097	SEMANTIC	8	0.0516	EITHER	4	0.0258	PRESENTED	9	0.0581

Table 5.7: Most frequent lexical items for abstracts in the AbstRA corpus

Nouns			Adjectives			Adverbs			Verbs		
Word	Freq	%	Word	Freq	%	Word	Freq	%	Word	Freq	%
TIME	777	0.2191	SUCH	633	0.1785	NOT	1264	0.3565	IS	6158	1.7368
NUMBER	644	0.1816	OTHER	591	0.1667	AS	775	0.2186	BE	2518	0.7102
PACKETS	570	0.1608	DIFFERENT	427	0.1204	THEN	690	0.1946	ARE	2473	0.6975
ANALYSIS	454	0.1280	SAME	358	0.1010	ALSO	659	0.1859	WAS	1390	0.3920
CASE	439	0.1238	HIGH	337	0.0950	ONLY	506	0.1427	WERE	1086	0.3063
PROBLEM	432	0.1218	FIRST	323	0.0911	HOWEVER	380	0.1072	HAS	871	0.2457
ALGORITHM	428	0.1207	MOST	284	0.0801	THEREFORE	351	0.0990	HAVE	764	0.2155
RESULTS	404	0.1139	LOW	280	0.0790	THUS	351	0.0990	BEEN	515	0.1453
MODEL	367	0.1035	LEAST	254	0.0716	SO	320	0.0903	USED	504	0.1421
TEMPERATURE	357	0.1007	LARGE	245	0.0691	MORE	290	0.0818	USING	481	0.1357
SET	342	0.0965	NEW	223	0.0629	NOW	199	0.0561	GIVEN	420	0.1185
CONDITIONS	321	0.0905	MORE	211	0.0595	WELL	198	0.0558	SHOWN	333	0.0939
PROCESS	315	0.0888	SIMILAR	206	0.0581	OUT	197	0.0556	SEE	299	0.0843
SIZE	306	0.0863	SMALL	205	0.0578	RESPECTIVELY	183	0.0516	HAVE	287	0.0809
SEQUENCE	304	0.0857	POSSIBLE	201	0.0567	MOST	182	0.0513	LET	285	0.0804
FUNCTION	301	0.0849	TOTAL	193	0.0544	VERY	180	0.0508	FOLLOWING	270	0.0762
PRIORITY	290	0.0818	EXPERIMENTAL	183	0.0516	HERE	178	0.0502	DOES	248	0.0699
VALUE	286	0.0807	OPTIMAL	183	0.0516	EVEN	169	0.0477	OBTAINED	239	0.0674
PACKET	278	0.0784	SINGLE	183	0.0516	HENCE	149	0.0420	FOLLOWS	205	0.0578
RATE	277	0.0781	MANY	172	0.0485	FIRST	143	0.0403	FOUND	205	0.0578

Table 5.8: Most frequent lexical items for RAs in the AbstRA corpus

Nouns			Adjectives			Adverbs			Verbs		
Word	Freq	%	Word	Freq	%	Word	Freq	%	Word	Freq	%
Computer Science											
PROBLEM	34	0.8843	FUZZY	12	0.3121	ALSO	11	0.2861	IS	60	1.5605
SPACE	23	0.5982	INTUITIONISTIC	9	0.2341	NOT	6	0.1560	ARE	31	0.8062
K	20	0.5202	LINEAR	9	0.2341	AS	5	0.1300	BE	20	0.5202
ALGORITHM	18	0.4681	CONSTANT	8	0.2081	IN	5	0.1300	SHOW	14	0.3641
NETWORK	16	0.4161	OPTIMAL	6	0.1560	ONLY	5	0.1300	HAVE	12	0.3121
GRAPHS	15	0.3901	DETERMINISTIC	5	0.1300	HOWEVER	4	0.1040	BEEN	11	0.2861
TIME	15	0.3901	METRIC	5	0.1300	RESPECTIVELY	4	0.1040	GIVEN	11	0.2861
MODEL	13	0.3381	POLYNOMIAL	5	0.1300	BEST	3	0.0780	USING	9	0.2341
APPROXIMATION	12	0.3121	REAL	5	0.1300	FINALLY	3	0.0780	HAS	7	0.1821
GRAPH	11	0.2861	ARBITRARY	4	0.1040	FURTHERMORE	3	0.0780	BASED	6	0.1560
Linguistics											
UNIT	10	0.4517	SEMANTIC	8	0.3613	NOT	8	0.3613	IS	40	1.8067
ANALYSIS	9	0.4065	INTONATIONAL	5	0.2258	ALSO	5	0.2258	ARE	18	0.8130
THEME	8	0.3613	OTHER	5	0.2258	ONLY	3	0.1355	BE	8	0.3613
SENTENCE	7	0.3162	SYNTACTIC	5	0.2258	MORE	3	0.1355	ARGUE	4	0.1807
ARTICLE	6	0.2710	DIFFERENT	4	0.1807	AS	3	0.1355	BASED	4	0.1807
RESULTS	6	0.2710	LEXICAL	4	0.1807	SEMANTICALLY	2	0.0903	MADE	4	0.1807
LANGUAGE	6	0.2710	PROPER	4	0.1807	PURELY	2	0.0903	WAS	4	0.1807
CONSTRUCTIONS	6	0.2710	THIRD	4	0.1807	POSSIBLY	2	0.0903	ACCOUNT	3	0.1355
ADDITION	5	0.2258	REFERENTIAL	3	0.1355	NATURALLY	2	0.0903	BOUNDED	3	0.1355
EXPERIMENT	5	0.2258	SECOND	3	0.1355	HOWEVER	2	0.0903	GIVEN	3	0.1355

Table 5.9: Most frequent lexical items for abstracts in the disciplines of computer science and linguistics

Nouns			Adjectives			Adverbs			Verbs		
Word	Freq	%	Word	Freq	%	Word	Freq	%	Word	Freq	%
Biology											
GENE	24	0.4409	MITOCHONDRIAL	15	0.2755	ALSO	10	0.1837	WAS	28	0.5143
PROTEINS	24	0.4409	HUMAN	10	0.1837	HERE	9	0.1653	ARE	21	0.3857
PROTEIN	20	0.3674	GENOMIC	8	0.1470	HOWEVER	7	0.1286	WERE	19	0.3490
DNA	18	0.3306	OTHER	8	0.1470	ONLY	6	0.1102	HAVE	13	0.2388
CELLS	16	0.2939	PHI	8	0.1470	AS	5	0.0918	BE	9	0.1653
METHYLATION	14	0.2572	DIFFERENT	7	0.1286	VERY	5	0.0918	BEEN	9	0.1653
SEQUENCE	14	0.2572	FUNCTIONAL	6	0.1102	EXCLUSIVELY	3	0.0551	CONTAINING	9	0.1653
ANALYSIS	12	0.2204	NEW	6	0.1102	POTENTIALLY	3	0.0551	HAS	9	0.1653
STRUCTURE	12	0.2204	BINDING	5	0.0918	THUS	3	0.0551	FOUND	7	0.1286
GC	10	0.1837	EXTENDED	5	0.0918	TYPICALLY	3	0.0551	REVEALED	7	0.1286
Mechanical engineering											
PROCESS	24	0.6012	EXPERIMENTAL	10	0.2505	AS	4	0.1002	IS	63	1.5782
TEMPERATURE	22	0.5511	DIFFERENT	7	0.1754	WELL	4	0.1002	ARE	29	0.7265
HEAT	20	0.5010	HIGH	7	0.1754	ALSO	3	0.0752	WAS	29	0.7265
CONCENTRATION	14	0.3507	LIQUID	7	0.1754	HOWEVER	3	0.0752	WERE	15	0.3758
TRANSFER	14	0.3507	CONSTANT	6	0.1503	OFTEN	3	0.0752	BEEN	13	0.3257
RATE	13	0.3257	THERMAL	6	0.1503	VERY	3	0.0752	BASED	11	0.2756
TIME	13	0.3257	LINEAR	5	0.1253	CONTINUOUSLY	2	0.0501	BE	11	0.2756
FLOW	12	0.3006	NUMERICAL	5	0.1253	EXPERIMENTALLY	2	0.0501	HAS	9	0.2255
AIR	11	0.2756	SUPERFICIAL	5	0.1253	EXTERNALLY	2	0.0501	HAVE	8	0.2004

Table 5.10: Most frequent lexical items for abstracts in the disciplines of biology and mechanical engineering

Nouns			Adjectives			Adverbs			Verbs		
Word	Freq	%	Word	Freq	%	Word	Freq	%	Word	Freq	%
Computer Science											
PACKETS	570	0.5040	SUCH	258	0.2281	THEN	471	0.4164	IS	2574	2.2758
TIME	451	0.3987	MOST	224	0.1980	NOT	353	0.3121	BE	1015	0.8974
V	434	0.3837	LOW	178	0.1574	AS	248	0.2193	ARE	796	0.7038
NUMBER	425	0.3758	LEAST	174	0.1538	ALSO	198	0.1751	HAS	354	0.3130
ALGORITHM	411	0.3634	HIGH	163	0.1441	THEREFORE	177	0.1565	HAVE	289	0.2555
G	332	0.2935	OPTIMAL	162	0.1432	SO	165	0.1459	LET	267	0.2361
PROBLEM	329	0.2909	OTHER	157	0.1388	THUS	157	0.1388	GIVEN	209	0.1848
K	292	0.2582	FIRST	156	0.1379	ONLY	155	0.1370	FOLLOWS	170	0.1503
PRIORITY	290	0.2564	SAME	136	0.1202	NOW	125	0.1105	USING	154	0.1362
PACKET	277	0.2449	LARGE	133	0.1176	HENCE	102	0.0902	FOLLOWING	151	0.1335
Linguistics											
CONTEXT	169	0.1566	S	156	0.1445	MORE	131	0.1214	IS	1861	1.7242
SENTENCES	160	0.1482	SEMANTIC	149	0.1381	HOWEVER	110	0.1019	ARE	846	0.7838
NAMES	144	0.1334	DIFFERENT	123	0.1140	THEN	108	0.1001	BE	731	0.6773
INFORMATION	141	0.1306	PROPER	121	0.1121	SO	106	0.0982	WAS	379	0.3512
CLAUSE	135	0.1251	SAME	108	0.1001	OUT	93	0.0862	WERE	316	0.2928
SENTENCE	132	0.1223	SYNTACTIC	103	0.0954	RATHER	93	0.0862	HAS	255	0.2363
CASE	130	0.1204	FIRST	102	0.0945	THUS	90	0.0834	HAVE	221	0.2048
WH	130	0.1204	DEFINITIVE	98	0.0908	EVEN	85	0.0788	BEEN	151	0.1399
CONDITIONS	127	0.1177	MANY	93	0.0862	HERE	76	0.0704	USED	148	0.1371
THEME	119	0.1103	NEW	93	0.0862	THEREFORE	73	0.0676	SEE	143	0.1325

Table 5.11: Most frequent lexical items for RAs in the disciplines of computer science and linguistics

Nouns			Adjectives			Adverbs			Verbs		
Word	Freq	%	Word	Freq	%	Word	Freq	%	Word	Freq	%
Biology											
AL	410	0.6221	DIFFERENT	102	0.1548	NOT	188	0.2853	IS	578	0.8771
ET	371	0.5630	OTHER	101	0.1533	ALSO	121	0.1836	WAS	546	0.8285
PROTEINS	271	0.4112	MITOCHONDRIAL	91	0.1381	AS	107	0.1624	WERE	489	0.7420
PROTEIN	226	0.3429	HUMAN	81	0.1229	ONLY	86	0.1305	ARE	352	0.5341
GENE	192	0.2913	SIMILAR	70	0.1062	HOWEVER	82	0.1244	BE	249	0.3778
DNA	161	0.2443	SAME	55	0.0835	THUS	57	0.0865	USING	195	0.2959
SEQUENCE	157	0.2382	SUCH	55	0.0835	PREVIOUSLY	55	0.0835	HAVE	154	0.2337
C	150	0.2276	BINDING	52	0.0789	THEREFORE	47	0.0713	BEEN	135	0.2049
GC	149	0.2261	SINGLE	48	0.0728	MORE	44	0.0668	HAS	127	0.1927
CELLS	143	0.2170	HIGH	47	0.0713	WELL	40	0.0607	USED	115	0.1745
GENES	141	0.2140	GENOMIC	43	0.0652	THEN	38	0.0577	SHOWN	85	0.1290
Mechanical engineering											
TEMPERATURE	308	0.4555	DIFFERENT	108	0.1597	NOT	119	0.1760	IS	1145	1.6932
BED	257	0.3800	SUCH	91	0.1346	AS	118	0.1745	BE	523	0.7734
HEAT	250	0.3697	LIQUID	90	0.1331	ALSO	100	0.1479	ARE	479	0.7083
FLOW	212	0.3135	EXPERIMENTAL	86	0.1272	HOWEVER	91	0.1346	WAS	390	0.5767
FIG	211	0.3120	HIGH	81	0.1198	ONLY	75	0.1109	WERE	256	0.3786
TIME	210	0.3105	CONSTANT	74	0.1094	THEN	73	0.1079	HAS	135	0.1996
PROCESS	205	0.3031	THERMAL	74	0.1094	RESPECTIVELY	57	0.0843	USED	133	0.1967
VELOCITY	191	0.2824	OTHER	68	0.1006	THEREFORE	54	0.0799	SHOWN	121	0.1789
RATE	183	0.2706	DUE	67	0.0991	WELL	48	0.0710	BEEN	117	0.1730
CONCENTRATION	181	0.2677	SAME	59	0.0872	THUS	47	0.0695	OBTAINED	112	0.1656

Table 5.12: Most frequent lexical items for RAs in the disciplines of biology and mechanical engineering

5.1.2.3 Keywords

The previous section, although explorative due to the size of the AbstRA corpus, discussed vocabulary differences between abstracts and RAs as well as lexical variety across disciplines. However, frequencies of occurrence of words in corpora *per se* do not give any information whether high frequencies of a given word are to be interpreted as particularly characteristic of the particular corpora under study or whether such a high frequency would conform with the expectations for general English. In order to make such inferences, a comparison of the results obtained for the corpus under study with a reference corpus of general English is needed. Through a comparison, it can be noticed that the frequency of occurrence of some words in the corpus under study are unexpectedly high in comparison to the frequency of occurrence of the same word in the reference corpus. Such words are called *keywords*. A formal definition is given by Scott (2008): “Keywords are those whose frequency is unusually high in comparison with some norm”. The corresponding quantitative evaluation of keywords is called *keyness*. It “relates to the frequency of particular lexical items within a text as compared with their frequency in a reference corpus” Scott (2001: 109). A detailed discussion on the nature of keyness itself and on keyness in specific discourse contexts is found in Bondi & Scott (2010).

An analysis of keywords is performed with the help of the WordSmith Tools. First, a comparison of a word list with a word list of a reference corpus of texts is performed. Then, the tool examines “each word-form and compares its frequency as a percentage of the text with the frequency of the same word-form in the reference” (Scott 2008). Some of the words are outstandingly more frequent in the corpus under study and are therefore marked as keywords[44]. Keywords typically reflect characteristics of *aboutness* and *style* of the corpus under study.

This section thus reports on a keyword analysis of the AbstRA corpus and the BNC[45] as reference corpus. However, it should be noted that the size of the AbstRA corpus is smaller than the recommended size of 500 texts (Scott 2008). Nevertheless, the results provide some insights on the keywords in the AbstRA corpus and support the discussion on its results

[44] Threshold: p-value = 0.000001. (`http://www.lexically.net/downloads/version5/HTML/?keywords_calculate_info.htm` (accessed: 14 October 2010)).

[45] “The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written” `http://www.natcorp.ox.ac.uk/` (accessed: 21 Oktober 2010).

Keyword – Abstracts	Freq.	%	BNC Freq.	BNC %	Keyness	p-value
PROBLEM	39	0.2540	28576	0.0287	100.84	2.67848E-15
ANALYSIS	30	0.1954	13130	0.0132	105.72	2.1684E-15
SPACE	35	0.2279	12601	0.0127	136.17	7.50176E-16
MODEL	28	0.1823	13155	0.0132	94.98	3.52763E-15
RESULTS	28	0.1823	15337	0.0154	87.07	5.35263E-15
PROCESS	26	0.1693	22499	0.0226	59.63	4.75174E-14
GENE	25	0.1628	2231		164.69	3.58367E-16
PROTEINS	24	0.1563	1262		183.04	2.41666E-16
TEMPERATURE	26	0.1693	4343		139.44	6.82402E-16
HEAT	25	0.1628	5794		118.24	1.33803E-15
WE	101	0.6577	300833	0.3025	48.00	2.91513E-13
Keyword – Research articles	Freq.	%	BNC Freq.	BNC %	Keyness	p-value
NUMBER	643	0.1829	48885	0.0491	745.44	2.22169E-18
ANALYSIS	460	0.1308	13130	0.0132	1270.53	4.23133E-19
CASE	463	0.1317	45216	0.0455	376.73	1.98715E-17
PROBLEM	440	0.1251	28576	0.0287	613.23	4.11655E-18
ALGORITHM	462	0.1314	552		3826.00	1.46718E-20
RESULTS	414	0.1177	15337	0.0154	955.63	1.02193E-18
MODEL	404	0.1149	13155	0.0132	1022.54	8.28158E-19
TEMPERATURE	391	0.1112	4343		1749.23	1.58533E-19
SET	492	0.1399	44247	0.0445	454.21	1.07819E-17
CONDITIONS	322	0.0916	15376	0.0155	605.95	4.27575E-18
WE	2396	0.6814	300833	0.3025	1226.35	4.71928E-19
Threshold: p-value = 0.000001						

Table 5.13: Keyness of some frequent item in the AbstRA corpus in comparison to BNC

for the frequency of occurrence of lexical items in the previous section (cf. Section 5.1.2.2). It is not the purpose of this analysis to evaluate keywords and keyness in the AbstRA corpus throughly because of the restrictions of size. There are however some interesting results to be reported. Table 5.13 shows the frequencies of occurrence of some of the most frequent words in abstracts and in RAs (cf. Tables 5.7 and 5.8) in comparison to their frequencies of occurrence in the reference corpus, BNC, the corresponding keyness and its p-value. First, it can be observed that the most frequent words in abstracts and RAs are not as frequent in the reference corpora. For example, the frequency of occurrence of *problem* in abstracts is 0.2540% and 0.1251% in RAs, but only 0.0287% in the BNC. This indicates that *problem* is a much more relevant topic in abstracts than in RAs and in the reference corpus, which is supposed to represent general English. From the values of frequencies of occurrence, WordSmith calculates the keyness between the corpora under study and the reference corpus together with its respective p-value. For the case of *problem* the keyness values are 100.84 for abstracts and 613.23 for RAs, both p-values are significant since they are smaller than 0.05. The keyness value of 100.84 for *problem* in abstracts indicates that *problem* occurs a bit more than one hundred times in the abstract corpus as in the BNC. Similarly, the word *problem* is more than 600% more frequent in RAs than in the BNC. Such results support the interpretation of the data of most frequent lexical items as an indication of domain specificity (cf. Section 5.1.2.2). This is what was meant by keyness being an indication of the "aboutness" of texts.

Furthermore, keyness can also provide some insights on style characteristics of texts. Although this research does not to deal with style issues, it is interesting to observe that there are clear style differences between the AbstRA corpus and the BNC as reference corpus. One indicator of such differences is the frequency of occurrence of the personal pronoun *we*. According to the results of Table 5.13, *we* has a relative frequency of occurrence of 0.6577% in abstracts, 0.6814% in RAs and 0.3025% in the BNC. This data together with the corresponding keyness values for *we* indicate that in abstracts and, most of all, in RAs *we* is proportionally much more frequently used as in general English, therefore marking style in both text types. According to Conrad & Biber (2001: 88), the use of *we* is an act of persuasion because it demands the participation of readers in authors' perspectives on a given issue [46].

The lexical features presented here, although exploratory, revealed qualitative and quantitative differences between abstracts and RAs as well as between each of these corpora and the reference corpus of general English, also across disciplines. Abstracts and RAs employ the spectrum of possible lexical realization quite differently as compared to general English, specially concerning the use of domain specific lexical items. Complementarily to this section, Section 5.1.3, discusses the results of the quantitative analysis of the grammatical features chosen for the empirical analysis in this research.

[46] The same analysis procedure could have been made for the most frequent lexical items across disciplines, to reveal domain specific characteristics of the corpus under study in comparison to general English. However, due to the size of the AbstRA corpus this analysis was not performed (see page 101 for the discussion on the size of the corpus and lexical analyis.

5.1.3 Grammatical features

This section reports on the results concerning the grammatical features chosen for empirical analysis of the AbstRA corpus. First, Section 5.1.3.1 discusses the distribution of modals auxiliaries in the AbstRA corpus. Section 5.1.3.2 then presents the results of the use of passive voice, while Section 5.1.3.3 discusses the use nominalizations in abstracts and RAs, also across disciplines. Finally, the issue of grammatical intricacy in the corpus under study is addressed in Section 5.1.3.4.

5.1.3.1 Modals

Modals in language are used mostly for marking persuasion, i.e., they mark the author's "own assessment of likelihood or advisability" (Biber 1988: 148). Modals can be classified into three categories according to their function in language: *predictive* modals for referring to the future, *possibility* modals, which are used to linguistically realize different perspectives on a topic; and *necessity* modals, which directly express persuasiveness. In English, this classification can be applied to modals, according to Biber (1988: 241ff.), as follows:

- Possibility modals: *can, may, might, could*
- Necessity modals: *ought, should, must*
- Predictive modals: *will, would, shall*

Examples 5.20 to 5.28 illustrate the occurrence of modals in the AbstRA corpus.

(5.20) This concept *can* enable one to conduct continuous productions of CNCs and MWCNTs. [abstract.C3.7; emphasis added]

(5.21) Pinch analysis improves energy efficiency for batch processes and it may increase the productivity of a revised plant. [abstract.C3.13; emphasis added]

(5.22) On the other hand, the material *might* be difficult to fluidise. [abstract.C3.6; emphasis added]

(5.23) [...] they cannot themselves form part of an innate Universal Grammar, and neither *could* they be interpreted in terms of simple parameter setting. [RA.C1.12; emphasis added]

(5.24) [. . .] which implies that in this case a competitive ratio of *should* not be considered impressive. [RA.A.14; emphasis added]

(5.25) The ideal resource assignment *must* balance the utilization of the underlying system against the loss of [. . .] among several servers. [abstract.A.9; emphasis added]

(5.26) This paper *will* explore the Theme unit as it functions to organise discourse in Japanese [. . .] [abstract.C1.9; emphasis added]

(5.27) [. . .] DT40 clones *would* likely contain sequences with large numbers of unique GCs [. . .] [RA.C2.13; emphasis added]

(5.28) We *shall* consider a fluid model that is the limit as the packet sizes and the burst parameters [. . .] tend to zero. [RA.A.26; emphasis added]

Due to their persuasive character, modals are expected to be used more in argumentation than in exposition texts. Since abstracts are supposed to summarize RAs in very few words and limited space, it can be assumed that abstracts tend to be more expository than RAs and tend to use modals proportionally less than RAs. Thus, the null hypothesis to be tested, H_0, and its counterpart, the alternative hypothesis H_1, can be formulated as follows:

H_1: Abstracts show significantly lower frequency of occurrence of modals in comparison to their RAs.

H_0: Abstracts *do not* show significantly lower frequency of occurrence of modals in comparison to their RAs.

The quantification of modals was performed with WordSmith Tools, searching for the strings *will, would, shall, can, may, might, could, ought, should, must* over the AbstRA corpus. There is a small difference between the total number of modals per discipline and per text type (abstracts or RAs), respectively, in comparison to the parts-of-speech tag MD (modals) shown in Table 5.1. This small difference is due to the part-of-speech tagger, which is not free of errors (cf. Section 4.2). However this does not influence the results presented in this section.

Table 5.14 presents the results for modals in abstracts and RAs, while the frequency of occurrence of modals per discipline is shown in Table 5.15. As known from Table 5.1 in Section 5.1.1.1, the relative frequency of occurrence of modals in abstracts is 0.33% and in RAs 0.77%. This is a strong indication that RAs are much more argumentative than abstracts, which would be then more expository as expected.

Modal		Abstracts	RAs
possibility-modal			
	can	31	1288
	may	21	417
	might	0	108
	could	2	230
necessity-modal			
	ought	0	0
	should	4	204
	must	3	184
predictive-modal			
	will	6	542
	would	0	290
	shall	0	39

Fisher's Exact Test for Count Data, p-value = 0.0002374
χ^2 = 31.3454, df = 8, p-value = 0.0001219

	Abstracts	RAs
Σ possibility-modals	54	2043
Σ necessity-modals	7	388
Σ predictive-modals	6	871

Fisher's Exact Test for Count Data, p-value = 0.001491

Table 5.14: Frequency of occurrence of modals in the AbstRA corpus

According to Table 5.14, the first interesting observation is that *ought* does not occur at all in the AbstRA corpus. *Ought* can be used to express obligation, duty, or necessity. However, it seems that for the AbstRA corpus these properties are fully covered by the use of *must*. For both abstracts and RAs, the most frequent type of modals are those expressing possibility, followed by necessity and then prediction modals. The results are tested for significance in two different ways. First, the raw frequencies of occurrence of *can, may, might, could, should, must, will, would* and *shall* for abstracts and RAs are tested for significance with the Fisher's test. The Fisher's test calculates the p-values exactly, while the chi-square test always depends on approximations. Therefore, the Fisher's test should be always the first choice of test for significance of such data. The respective p-value of 0.0002374 indicates that there is a significant difference between the distribution of modals in abstracts in comparison to RAs. Furthermore, when analyzed in terms of categories of modals, i.e., possibility, necessity, predictive, the corresponding result of the Fisher's test is p-value = 0.001491.

Modal	Abstracts								RAs							
	Computer science		Linguistics		Biology		Mechanical engineering		Computer science		Linguistics		Biology		Mechanical engineering	
	F	%	F	%	F	%	F	%	F	%	F	%	F	%	F	%
can	13	65.00	3	21.43	3	18.75	12	70.59	618	49.05	269	25.19	84	23.01	317	52.05
may	3	15.00	7	50.00	6	37.50	5	29.41	91	7.22	182	17.04	72	19.73	72	11.82
might	0	0.00	0	0.00	0	0.00	0	0.00	20	1.59	50	4.68	29	7.95	9	1.48
could	0	0.00	0	0.00	2	12.50	0	0.00	39	3.10	63	5.90	85	23.29	43	7.06
ought	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
should	1	5.00	0	0.00	3	18.75	0	0.00	36	2.86	105	9.83	21	5.75	42	6.90
must	2	10.00	1	7.14	0	0.00	0	0.00	87	6.90	69	6.46	1	0.27	27	4.43
will	1	5.00	3	21.43	2	12.50	0	0.00	288	22.86	154	14.42	32	8.77	68	11.17
would	0	0.00	0	0.00	0	0.00	0	0.00	62	4.92	166	15.54	41	11.23	21	3.45
shall	0	0.00	0	0.00	0	0.00	0	0.00	19	1.51	10	0.94	0	0.00	10	1.64
Σ possibility-modals	16	80.00	10	71.43	11	68.75	17	100.00	768	60.95	564	52.81	270	73.97	441	72.41
Σ necessity-modals	3	15.00	1	7.14	3	18.75	0	0.00	123	9.76	174	16.29	22	6.03	69	11.33
Σ predictive-modals	1	5.00	3	21.43	2	12.50	0	0.00	369	29.29	330	30.90	73	20.00	99	16.26
Σ modals	20	100.00	14	100.00	16	100.00	17	100.00	1260	100.00	1068	100.00	365	100.00	609	100.00

Table 5.15: Frequency of occurrence of modals across disciplines in the AbstRA corpus

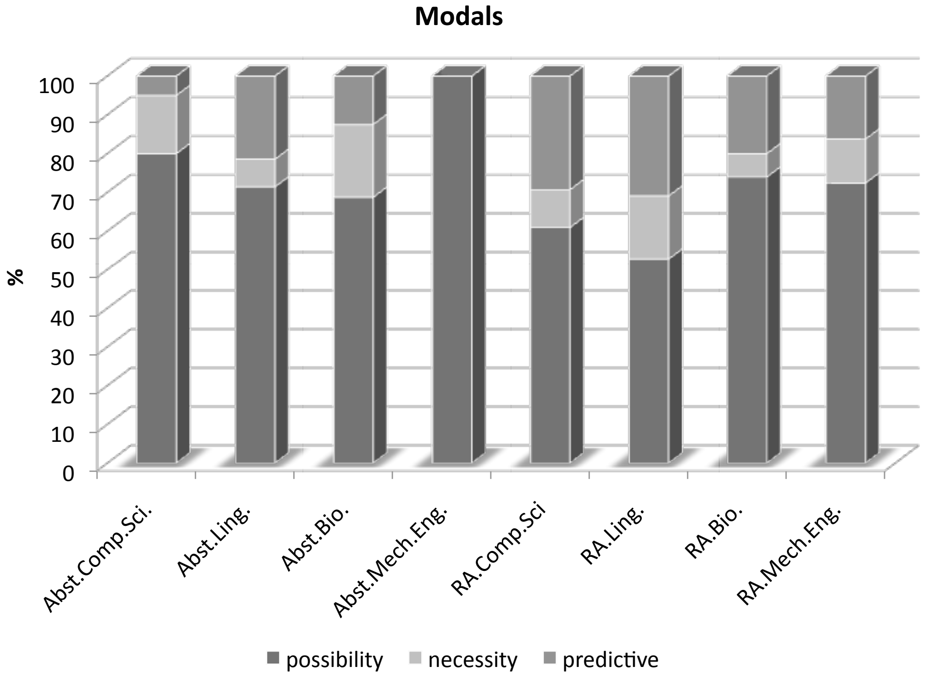


Figure 5.13: Modals in the AbstRA corpus

Hence, these results allow $\mathbf{H}_0$ to be refuted.

The variation on the use of modals across disciplines is shown in Table 5.15 and in Figure 5.13. Apart from the fact that modals are not very often used in abstracts, there are some interesting observations to be made. Possibilities modals are over all disciplines the most frequent kind of modals used in abstracts. Necessity modals occur mainly in abstracts of computer science and biology, followed by linguistics. While abstracts from mechanical engineering make no use of necessity or predictive modals, predictive modals are very frequent in abstracts of linguistics, followed by biology. Thus, the data indicate that, while all disciplines rely on the use of modals for addressing different perspectives on a given issue (i.e., through possibility modals), computer science makes very frequent use of direct persuasiveness (i.e., through necessity modals - 15.00%). Linguistics tends to refer more to future possibilities (i.e., through predictive modals - 21.43%), and biology shows a more balanced use of modals.

Research articles show a slightly different profile on the use of modals across disciplines (cf. Table 5.15 and Figure 5.13). Interestingly, RAs from the discipline of linguistics are the ones using the fewest number of possibility modals (52.81%). Biology and computer science make very little use of necessity modals with 9.76% and 6.03%, respectively. Finally, the frequency of occurrence of predictive modals are relatively similar for RAs across disciplines. However, predictive modals are almost twice as frequent in linguistics (30.90%) than in mechanical engineering (16.26%).

5.1.3.2 Passives

Passives are acknowledged to be a typical characteristic of scientific discourse (cf. Banks 2008; Gustafsson 2006; Halliday & Martin 1993). Passives are used mostly when the role of the agent of an action is not that important and they characterize objectiveness in discourse. According to Biber (1988: 228), "[i]n passive constructions, the agent is demoted or dropped altogether, resulting in a static, more abstract presentation of information".

> Even more importantly, passive voice allows concepts and objects (rather than people) to be the grammatical subject of the sentence, making the discourse topic clear. (Biber & Conrad 2009: 123)

Since abstracts are supposed to summarize RAs and condense information, it can be assumed that abstracts make use of passive constructions more frequently than RAs. Thus, the null hypothesis to be tested, H_0, and its counterpart, the alternative hypothesis H_1, can be formulated as follows:

H_1: Abstracts show significantly higher frequency of occurrence of passive constructions in comparison to their RAs.

H_0: Abstracts *do not* show significantly higher frequency of occurrence of passive constructions in comparison to their RAs.

The identification and extraction of passives is performed based on part-of-speech tagging using IMS/CWB for querying. Appendix A.4 describes all queries used for the extraction of passives in the AbstRA corpus. After the extraction, the instances of passive voice are classified according to *tense*, *aspect*, and *mood*, based on the criteria suggested by Gustafsson (2006), who investigated the development of passive in nineteenth-century scientific writing. Examples 5.29 to 5.37 show the classification criteria:

- Tense
  - Present (e.g., *may be chosen, is shown, are required, to be solved*)

    (5.29) Then the action *may be chosen* arbitrarily by the module, [. . .] [RA.A.23; emphasis added]
  - Past (e.g., *were achieved, have been developed, has been proposed*)

    (5.30) Other approaches *have been developed* to improve on the previous methods [. . .] [RA.C3.13; emphasis added]
  - Future (e.g., *will be presented, will be encoded*)

    (5.31) Justification for the theorisation of this textual unit *will be presented* together with a number of examples. [abstract.C1.9; emphasis added]
- Aspect
  - Indefinite (e.g., *is given, are required*)

    (5.32) We conclude that [. . .] a mitochondrially encoded gene product *is required* for promoting [. . .] [abstract.C2.22; emphasis added]
  - Perfect (e.g., *have been treated, has been developed*)

    (5.33) Whether a given graph [. . .], for which fast sequential and parallel algorithms *have been developed* in a sequence of papers. [abstract.A.27; emphasis added]
  - Progressive (e.g., *are being resolved*)

    (5.34) Based on this, [. . .] these AID-induced DSBs *are being resolved* either by HR or NHEJ. [RA.C2.13; emphasis added]
- Mood
  - Indicative (e.g., *is shown, are given, were developed*)

    (5.35) It *is shown* in this work that such a strategy [. . .] [abstract.C3.12; emphasis added]
  - Conditional (e.g., *may be approached, should be performed*)

    (5.36) For arbitrary metric spaces, this goal *may be approached* by using probabilistic metric approximation techniques. [abstract.A.7; emphasis added]
  - Imperative (e.g., *Let . . . be chosen*)

    (5.37) *Let* Q *be* an undirected tree *rooted* at s with leaf set L and depth h.. [RA.A.11; emphasis added]

	Abstracts		Research articles	
	F	%	F	%
Passive	323	55.69	7048	45.64
Non-passive	257	44.31	8393	54.36
Sentences	580	100.00	15441	100.00
Fisher's Exact Test for Count Data, p-value = 2.321e-06				

Table 5.16: Frequency of occurrence of passive constructions in the AbstRA corpus

The frequency of occurrence of passives is not only determined as raw values, but also as a relative frequency of a ratio between passives and non-passive, i.e., active or middle, sentences. For this reason, it is necessary to obtain the total number of sentences in abstracts and in RAs. Sentence boundary information is automatically generated by AnnoLab and TreeTagger, respectively, in the corpus processing steps and encoded in the corpus as annotation (cf. Section 4.2). This information and therefore the total number of sentences in each sub-corpus can be queried using IMS/CWB using the following syntax, as for example in the discipline of computer science, which is internally identified by the letter A:

```
Total_sentences = <s>[]*</s> :: match.document_uri = ".*/A/.*";
```

The results are given in terms of passive and *non*-passive constructions. The purpose of using the term *non*-passive constructions instead of active constructions, which can be seen as an approximation, is not to induce the false idea that a construction that is not passive is obligatorily in active voice in English. In the case of scientific writing, there are sentences without verbs, e.g., headlines, sometimes title, and those constructions which are not specifically in passive voice that are classified as *non*-passive.

Since the query on the total number of sentences was possible at a sub-corpus basis, the queries for passives were performed over all texts of a given discipline simultaneously. For this reason, the results correspond to overall abstracts, overall RAs and their overalls in each of the sub-corpus as a whole.

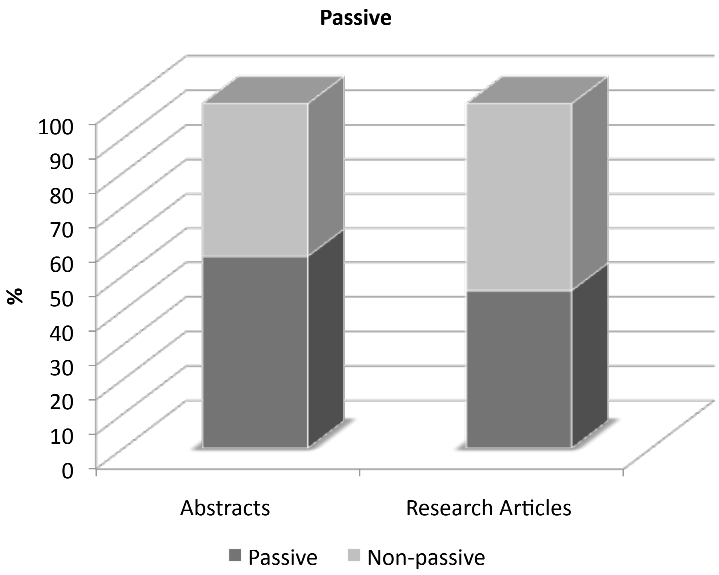


Figure 5.14: Passives in the AbstRA corpus

Table 5.16 shows the results of the number of passive and non-passive constructions in abstracts and in RAs in the AbstRA corpus, as well as the total number of sentences. These results are also displayed in Figure 5.14 for better visualization. Passive constructions are relatively more frequent in abstracts (55.69%) than in RAs (45.64%). In order to test whether this difference is significant, the raw data is tested with the Fisher test. The Fisher test result shows a p-value = 2.321e-06, indicating that the difference is statistically significant. For this reason, $\mathbf{H}_0$ can be rejected. Abstracts do make use of passives constructions significantly more frequently than RAs.

The next step is to classify the passives according to tense, aspect, and mood, as shown in Table 5.17 and in Figure 5.15. The raw frequencies of passive occurrence according to these three parameters is tested for significance with a Fisher's test. According to the results from the Fisher's test for *tense*, also displayed in Figure 5.15, there is no significant difference between the frequency of occurrence of passives in abstracts as compared to RAs. Thus, abstracts and RAs make use of the several tenses in passive constructions in English very similarly, although at first sight one could think that the difference in the use of future passives could contribute for significancy. However, there is a statistically significant difference in *aspect*.

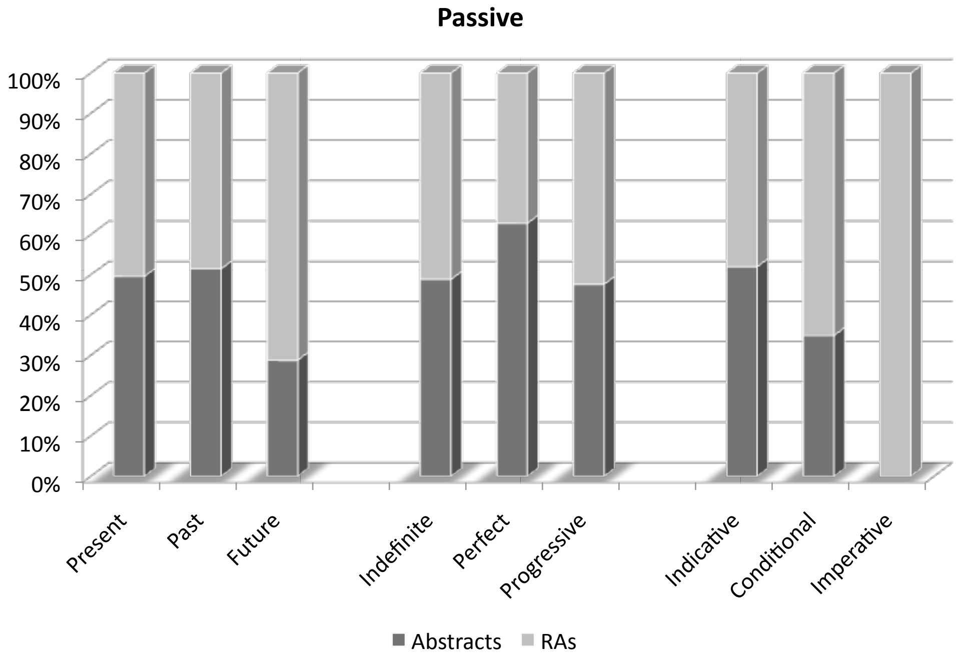


Figure 5.15: Passives in the AbstRA corpus

Passives	Abstracts		RAs	
	F	%	F	%
Tense				
Present	213	65.94	4727	67.07
Past	108	33.44	2213	31.40
Future	2	0.62	108	1.53
	Fisher's Exact Test for Count Data, p-value = 0.3673			
Aspect				
Indefinite	286	88.54	6537	92.75
Perfect	34	10.53	439	6.23
Progressive	3	0.93	72	1.02
	Fisher's Exact Test for Count Data, p-value = 0.01165			
Mood				
Indicative	296	91.64	5936	84.22
Conditional	27	8.36	1098	15.58
Imperative	0	0.00	14	0.20
	Fisher's Exact Test for Count Data, p-value = 0.0008357			

Table 5.17: Frequency of occurrence of passive constructions according to tense, aspect, and mood in the AbstRA corpus

Abstracts tend to use less indefinite, more perfect, and less progressive passives than RAs. Furthermore, their results according to *mood* are also statistically significant. While passives in abstracts and RAs occur mainly in the indicative mood, the relative occurrence of conditional passives in RAs is almost twice as in abstracts. This indicates a tendency to a more argumentative discourse in RAs. Moreover, imperative passives are only found in RAs and in a very low frequency of occurrence.

Finally, passive constructions in abstracts and RAs are classified according to their tense, aspect, and mood across disciplines. Table 5.18 shows the results of this classification in terms of raw frequency and the corresponding percentage. The results for *tense* show that while abstracts from computer science and linguistics mainly use passive constructions in the present tense, mechanical engineering and biology show a high frequency of occurrence in past tense passives, i.e., 37.21% and 50.62%, respectively. This observation is probably due to the typical procedural descriptions even in abstracts of such disciplines. A similar profile is also shown by RAs. However, RAs from biology use even more past (62.96%) than present (36.52%) passives. In contrast, RAs from computer science use mainly present passives (88.41%), followed by only 9.08% of past and 2.51% of future passives. This can be interpreted as an indication of very formal, abstract and mathematics-like discourse.

The analysis of the results for *aspect* reveals that no progressive passives are found in abstracts of biology and mechanical engineering. Furthermore, while computer science abstracts make frequently use of perfect passives (15.28%), linguistic abstracts use such passives the less, with on 2.44%. Research articles of the four disciplines show a very similar profile of the passive use according to aspect. Interestingly, however, is the fact that RAs in linguistics use progressive passives proportionally much more frequently in the other disciplines. This probably indicates a frequent work-in-progress report or similar on-going result report.

Furthermore, the data for passive construction according to *mood* indicate that no imperative passives occur in abstracts of any discipline. Moreover, abstracts of computer science use around twice the number of conditional passives than abstracts of the other disciplines. The mood classification of passives in RAs show that only imperative passives occur only in computer science. This is another indication for a very formal, mathematical calculating discourse, characteristic of this discipline. Furthermore, RAs from computer science are also the ones with the highest number of conditional passives, while biology is the discipline in which conditional passives are the less frequent.

The results of the analysis of passive voice in the AbstRA corpus can be summarized as follows: there is a highly significant difference in the frequency of occurrence of passives between abstracts and RAs, including a significant difference according to aspect and mood of passive constructions. Furthermore, differences across disciplines are observed, probably as a consequence of the different fields of discourse. Biology seems to be more experiential, i.e., they perform lots of actions and experiments and report about them, than the other disciplines, followed by mechanical engineering and then linguistics. Finally, computer science shows distinctive characteristics concerning the use of passive, as a consequence of a very formal and mathematical discourse.

Passives	Abstracts								RAs							
	Computer science		Linguistics		Biology		Mechanical engineering		Computer science		Linguistics		Biology		Mechanical engineering	
	F	%	F	%	F	%	F	%	F	%	F	%	F	%	F	%
Tense																
Present	58	80.56	34	82.93	40	49.38	81	62.79	1549	88.41	1415	73.58	562	36.52	1201	65.49
Past	13	18.06	6	14.63	41	50.62	48	37.21	159	9.08	481	25.01	969	62.96	604	32.93
Future	1	1.39	1	2.44	0	0.00	0	0.00	44	2.51	27	1.40	8	0.52	29	1.58
Σ	72	100.00	41	100.00	81	100.00	129	100.00	1752	100.00	1923	100.00	1539	100.00	1834	100.00
Aspect																
Indefinite	59	81.94	39	95.12	72	88.89	116	89.92	1653	94.35	1767	91.89	1405	91.29	1712	93.35
Perfect	11	15.28	1	2.44	9	11.11	13	10.08	89	5.08	112	5.82	126	8.19	112	6.11
Progressive	2	2.78	1	2.44	0	0.00	0	0.00	10	0.57	44	2.29	8	0.52	10	0.55
Σ	72	100.00	41	100.00	81	100.00	129	100.00	1752	100.00	1923	100.00	1539	100.00	1834	100.00
Mood																
Indicative	61	84.72	38	92.68	76	93.83	121	93.80	1364	77.85	1649	85.75	1430	92.92	1493	81.41
Conditional	11	15.28	3	7.32	5	6.17	8	6.20	374	21.35	274	14.25	109	7.08	341	18.59
Imperative	0	0.00	0	0.00	0	0.00	0	0.00	14	0.80	0	0.00	0	0.00	0	0.00
Σ	72	100.00	41	100.00	81	100.00	129	100.00	1752	100.00	1923	100.00	1539	100.00	1834	100.00
χ^2 = 45.1789, df = 21, p-value = 0.001641									χ^2 = 1405.065, df = 24, p-value < 2.2e-16							

Table 5.18: Frequency of occurrence of passive constructions across disciplines in the AbstRA corpus

5.1.3.3 Nominalizations

As previously mentioned, one of the most distinctive feature of abstracts is their information density. It is commonly known that complexity in scientific discourse is achieved mainly through specific terminology and nominalization, which is part of grammatical metaphor (cf. Halliday 2004a,b,c; Halliday & Martin 1993). Through nominalization, processes (linguistically realized as verbs) and properties (linguistically realized, in general, as adjectives) are re-construed metaphorically as nouns, enabling an informationally dense discourse. Through nominalization it is possible to build up chains or sequence of logical argument (Halliday 2008) and they are therefore “the single most powerful resource for creating grammatical metaphor” (Halliday 2004a: 656). For this reason, nominalization was chosen as a linguistic feature for this empirical analysis.

Due to their high informational density character, nominalizations are expected to be used more in abstracts than in RAs, which are supposed to condense the whole RA. Thus, the null hypothesis to be tested, H_0, and its counterpart, the alternative hypothesis H_1, can be formulated as follows:

H_1: Abstracts show significantly higher frequency of occurrence of nominalizations in comparison to their RAs.

H_0: Abstracts *do not* show significantly higher frequency of occurrence of nominalizations in comparison to their RAs.

The quantification of nominalizations was performed by WordSmith Tools, searching for words ending with the suffixes *-sion*, *-tion*, *-ment*, *-ness*, *-ity* and their plural forms in each single text over the AbstRA corpus. The choice of suffixes is based from the work of Biber (1988: 214). Examples 5.38 to 5.42 show instances of the corpus containing nominalizations.

(5.38) We do not demand n integers to be hashed into a table of size O(n) without any *collision*. [RA.A.19; emphasis added]

(5.39) A theoretical model for bubble breakup in slurry bubble columns as well as three-phase fluidized beds with fine particles has been developed based on an *exploration* into the *deformation*, *oscillation* and breakup process of the bubbles. [abstract.C3.16; emphasis added]

(5.40) The procedure, which enables to determine the basic transport properties of membrane/solution systems on the basis of the *measurement* in a continuous dialyzer at steady state, has been elaborated. [abstract.C3.1; emphasis added]

(5.41) However, the particular features proposed, such as the degree of *expectedness*, essentiality, and permanence, are not easily defined. [RA.C1.4; emphasis added]

(5.42) Finally, we investigate the *approximability* of several extensions of the load rebalancing model. [abstract.A.20; emphasis added]

The raw frequencies for nominalizations in the corpus are displayed in Table 5.19. It can be noticed that the number of instances of nominalizations per text varies considerably, depending not only on the discipline but probably also on the style used by the authors. The results are better visualized through a boxplot with notches, as shown in Figure 5.16. This figure shows a boxplot for the relative frequencies of nominalizations, i.e., raw frequencies divided per total number of tokens, which is calculated for each single text in the corpus.

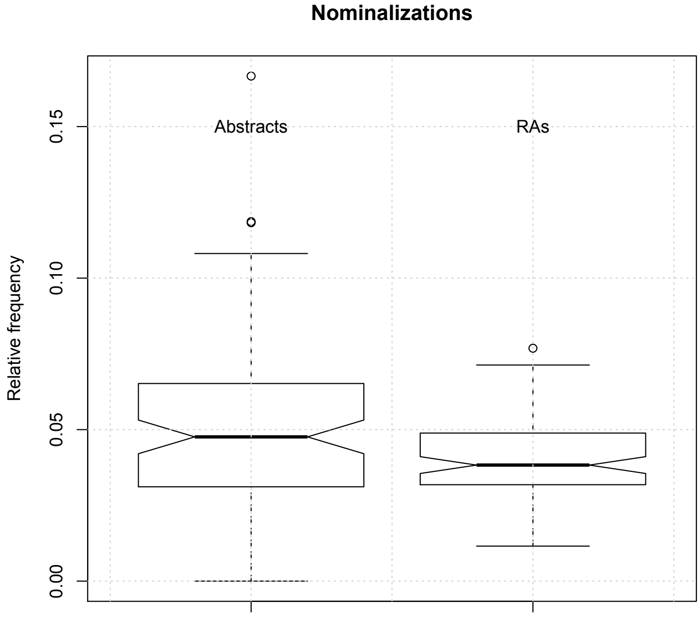


Figure 5.16: Nominalizations in the AbstRA corpus

According to this Figure, abstracts show a minimum relative frequency of nominalizations of 0.00, 1st quartile of 0.03118, median of 0.04760, mean of 0.05005, 3rd quartile of 0.06515, and a maximum relative frequency of

	Abstracts				Research articles			
	Computer science	Linguistics	Biology	Mechanical engineering	Computer science	Linguistics	Biology	Mechanical engineering
	2	13	6	9	46	230	131	168
	4	14	10	22	65	659	74	226
	13	14	2	15	114	473	124	139
	4	10	8	21	199	449	72	180
	3	13	7	6	40	83	141	123
	4	6	16	11	567	229	173	135
	6	7	15	11	112	386	184	170
	3	5	1	6	86	570	78	137
	12	17	8	12	114	431	87	184
	5	10	9	15	172	486	132	96
	1	5	10	6	112	323	170	74
	2	3	10	19	83	202	106	192
	10	13	12	5	334	554	180	102
	17	9	5	19	229	165	127	160
	2		16	10	86		108	170
	8		5	11	157		98	151
	0		7	12	37		138	84
	7		8	5	174		191	106
	9		7	8	194		79	169
	14		4	10	317		81	151
	7		14	9	139		148	209
	1		10	6	89		147	205
	2		8	3	123		118	87
	10		4	8	181		79	95
	3			12	69			81
	4			0	231			104
	10			2	119			43
				12				234
				8				101
Σ	163	139	202	293	4189	5240	2966	4076
Σ	797				16471			

Table 5.19: Nominalizations in the AbstRA corpus per text (raw frequencies)

nominalizations of 0.16670. Similarly, RAs show the following summary values: minimum relative frequency of nominalizations of 0.01155, 1st quartile of 0.03192, median of 0.03830, mean of 0.04052, 3rd quartile of 0.04877, and a maximum relative frequency of occurrence of nominalizations of 0.07685. The next step is the statistical evaluation of the data. An indication that the nominalization data are *not* normally distributed is that the values for median and mean are not identical. The Shapiro-Wilk test for normality testing is thus applied. According to a Shapiro-Wilk test, the distribution of the values for normality in abstracts deviate significantly from normality:

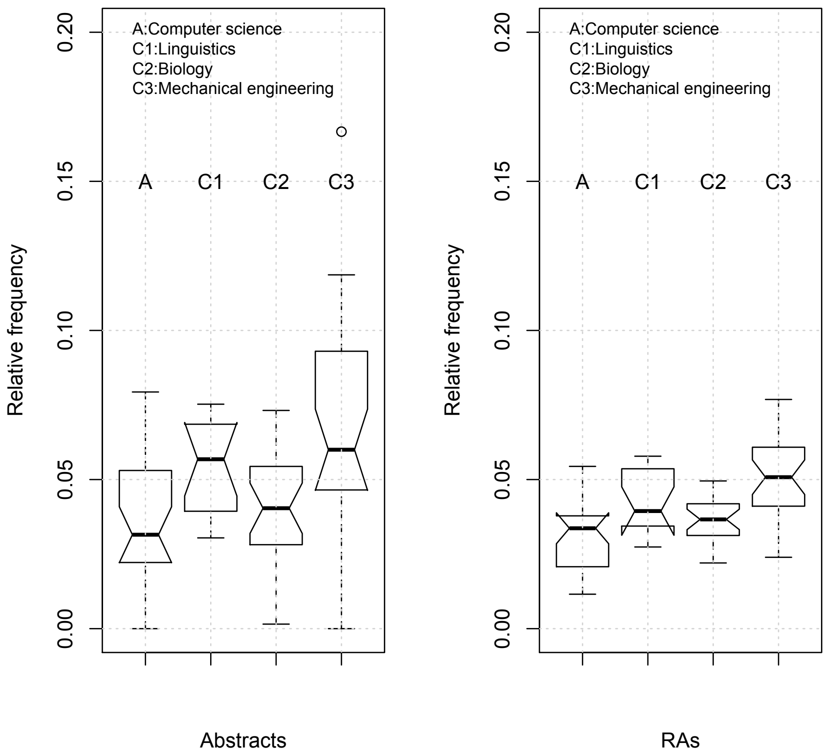


Figure 5.17: Nominalizations across disciplines in the AbstRA corpus

W = 0.9553, p-value = 0.002768. This is however not the case of RAs. The nominalizations' values are normally distributed in RAs since W = 0.9865, p-value = 0.4527.

Since the pre-requisite for normality in using a t-test was not met in the abstracts sub-corpus, a Wilcoxon rank-sum test must be applied anyway to test for significance. Because $\mathbf{H}_0$ is formulated as abstracts having significantly higher frequency of occurrence of nominalizations than their RAs, the one-tailed Wilcoxon rank-sum test in this direction is applied, i.e., in R with the parameter `alternative = "greater"`. The calculated value is W = 5328, p-value$_{one-tailored}$ = 0.007384. This means that there is a 99.26% of probability that this difference is *not* due to chance. Furthermore, $\mathbf{H}_{one-tailed}$ *can* be rejected because the p-value is higher than 0.05. Hence, abstracts show significantly higher frequency of occurrence of nominalizations in comparison to their RAs.

Suffix	Abstracts		Research articles	
	F	%	F	%
-sion	51	6.40	941	5.71
-sions	8	1.00	327	1.99
-tion	435	54.58	8544	51.87
-tions	89	11.17	2127	12.91
-ment	49	6.15	1269	7.70
-ments	28	3.51	544	3.30
-ness	13	1.63	245	1.49
-nesses	1	0.13	6	0.04
-ity	114	14.30	2295	13.93
-ities	9	1.13	173	1.05

Fisher's Exact Test for Count Data with simulated p-value (based on 5e+05 replicates), p-value = 0.1663

Table 5.20: Frequency of occurrence of nominalization suffixes in the AbstRA corpus

Again, the same procedure was followed for the analysis of nominalizations across disciplines. Figure 5.17 shows the boxplot with notches generated from the values in Table 5.19 for abstracts and RAs in each discipline, i.e., computer science (A), linguistics (C1), biology (C2), and mechanical engineering (C3). According to Figure 5.17, there is variation in the number of nominalizations between abstracts and RAs across all four disciplines. Abstracts from mechanical engineering present the highest relative frequency of nominalizations, while RAs from computer science show the lowest values.

When considering only abstracts and only RAs, it can be noticed that there is domain specific variation in the relative frequency of nominalizations since the boxes are vertically differently positioned. As for all shallow features, each pair of abstracts-RAs was tested for normality and significance. Again, most of the data is not normally distributed, which requires the use of a Wilcoxon rank-sum test for significance. The results from the Wilcoxon rank-sum test indicate that the relative frequency of normalizations for abstracts is significantly higher in comparison to their RAs only for the disciplines of linguistics (W = 145, $\text{p-value}_{one-tailed} = 0.01551$) and mechanical engineering (W = 562, p-value = 0.01378). For computer science and biology the results are not statistically significant since the obtained values for the Wilcoxon rank-sum test are W = 413, $\text{p-value}_{one-tailed} = 0.2047$ and W = 318, $\text{p-value}_{one-tailed} = 0.2715$, respectively.

Table 5.20 shows the distribution of nominalizations according to the individual suffixes, as raw frequencies and as percentage, in abstracts and RAs. Interestingly, the calculation of the Fisher-test in this case caused an overload in R, which was not able to complete this test. In such cases, it is recommended to adopt the chi-square test, unless it returns a warning concerning the no accuracy of the results. Since this is the case for this data, the next possible test for testing whether the differences in the distributions in Table 5.20 are significant is to perform a Fisher-test with simulated p-value and set a high number of replicates, in this case 5e+05, until the test is ended. This test thus returned a p-value = 0.1663, meaning that there is no statistically significant difference between these values since p-value > 0.05.

Therefore, it can be concluded that although the frequency of occurrence of nominalizations is statistically significantly higher in abstracts than in RAs, the forms of nominalizations chosen, i.e., the suffixes, and their proportions in abstracts are similar to those in RAs.

5.1.3.4 Grammatical complexity

Halliday & Martin (1993) argue that language is generally able to continuously adopt either a dynamic or a synoptic perspective. The dynamic perspective, also called the *doric* style, is characterized by a dynamic flow of happenings and processes. In contrast, the synoptic perspective, also called the *attic* style, represents language as a "world of things". The following sentences exemplify the doric and attic styles, respectively (Halliday & Martin 1993: 116):

- Attic
  e.g., *experimental emphasis becomes concentrated in testing the generalizations and consequences derived from the theories*
- Doric
  e.g., *we now start experimenting mainly in order to test whether things happen regularly as we would expect if we were explaining in the right way*

The attic style, which is very nominal and dense, developed much later in language, particularly through the development of scientific knowledge and the rise of scientific discourse.

> [T]he emergence of the new attic forms of expression added a new dimension to human experience: where previously there had been one mode of interpretation, the dynamic, now there were two, the synoptic and the dynamic – or rather, two poles, with varying degrees of semantic space possible between them.
>
> (Halliday & Martin 1993: 116)

Since abstracts are supposed to summarize and synthesize the knowledge of RAs, it is likely that abstracts make use of a more attic style than RAs. As already mentioned, the attic style is more nominalized than the doric one. In order to investigate the structural differences leading to an attic style, the frequency of occurrences of phrases reflecting a more nominalized style are quantitatively investigated in the parsed AbstRA corpus (cf. Section 4.2). Some of the features reflecting grammatical complexity and/or nominalized style are, among others, the number of nominal phrases (NP), prepositional phrases (PP), prepositional phrases which are embedded in nominal phrases (PP in NP), and finally, verbal phrases (VP). It is possible to query the frequency of occurrence of such phrases in the parsed AbstRA corpus using eXist and XQuery (cf. Section 4.2; p. 57 and 59). Examples of such queries are:

- Number of NPs
  `count(//Constituent[@cat="NP"])`
- Number of PPs
  `count(//Constituent[@cat="PP"])`
- Number of VPs
  `count(//Constituent[@cat="VP"])`
- Number of PPs embedded in NPs
  `count(//Constituent[@cat="PP" and ./parent::Constituent[@cat="NP"]])`
- Number of sentences
  `count(//Constituent[@cat="S"])`

It is not the aim of this section to thoroughly investigate the style characteristics of the texts under study. The queries presented above are just approximations that allow the researcher to gain insights into the attic and doric features of the AbstRA corpus. Following this approach, the null hypothesis to be tested, H_0, and its counterpart, the alternative hypothesis H_1, can be formulated as follows:

Phrase	Abstracts		Research articles	
	F	/S	F	/S
NP	1930	3.7622	56433	3.4450
PP	626	1.2203	17583	1.0734
PP in NP	317	0.6179	8027	0.4900
VP	791	1.5419	24557	1.4991
/S = per sentence				
Σ sentences	513		16381	
Fisher's Exact Test for Count Data, p-value = 0.02015				

Table 5.21: Frequency of occurrence of phrases in the AbstRA corpus

H$_1$: Abstracts show significantly higher frequency of occurrence of NPs, PPs, PPs embedded in NPs, and VPs per sentence in comparison to their RAs.

H$_0$: Abstracts *do not* show significantly higher frequency of occurrence of NPs, PPs, PPs embedded in NPs, and VPs per sentence in comparison to their RAs.

Table 5.21 displays the results of the queries for abstracts and RAs in the AbstRA corpus. The data is shown as raw frequency of occurrence (F) and its corresponding number per sentence, i.e., F / Σ sentences, so that the data are comparable. *All* selected phrases occur *more* frequently in abstracts than in RAs. In the corpus of abstracts, sentences have 3.76 nominal phrases on average, while RAs have 3.44 NPs/sentence. Prepositional phrases occur 1.22 times per sentence in abstracts and 1.073 in RAs. Prepositional phrases embedded in nominal phrases, which reflect the existence of large nominal groups and therefore compactness of information, occur 0.62 times per sentence in abstracts and 0.49 times per sentence in RAs. Finally, abstracts use in average 1.5 verbal phrases per sentence and RAs use in average 1.50 verbal phrases per sentence. In order to test whether these differences are statistically significant or not, the Fisher's exact test on the raw data was performed resulting in a p-value of 0.02015. Since this p-value is less than 0.05, **H**$_0$ can be rejected. Hence, the data indicate that abstracts show significantly higher frequency of occurrence of NPs, PPs, PPs embedded in NPs, and VPs, in comparison to their RAs. Thus, according to the data obtained from the AbstRA corpus, abstracts show a more attic style than their RAs. The same procedure was performed for each of the disciplines: computer science, linguistics, biology, and mechanical engineering.

The results of the frequency of occurrence of NPs, PPs, PPs in NPs, and VPs are shown in Table 5.22. The significance of the results for abstracts and for RAs is tested with a chi-square test because the number of data is too large for the Fisher's exact test. Both results are significant. Therefore, there is a significant difference between abstracts of the disciplines computer science, linguistics, biology, and mechanical engineering.

The same is valid for RAs; there is a significant difference in RAs across disciplines. Abstracts of mechanical engineering use the highest number of NPs/sentence, i.e., 4.24, while abstracts from linguistics only use 3.31 NPs/sentence. Prepositional phrases are more frequent in abstracts of mechanical engineering (1.62 PP/sentence) and less frequent in computer science (0.86 PP/sentence). Similarly, abstracts from mechanical engineering present the highest frequency of prepositional phrases embedded in nominal phrases, i.e., 0.86 PPinNP/sentence) and, again, computer science uses such construction the less, i.e., 0.45 PPinNP/sentence. The number of verb phrases is almost equally distributed across abstracts of these three disciplines. Thus, it can be said that abstracts of mechanical engineering show the most attic style in comparison to the abstracts of the other disciplines. Contrastively, it seems that abstracts of computer science use the less attic style of all the studied disciplines, probably as a consequence of a very formal discipline and consequently domain specific discourse.

The data for RAs across disciplines corroborates the argument that computer science is the less attic discipline. Meanwhile, RAs from mechanical engineering, followed by biology, show a higher frequency of the chosen phrases per sentence.

Phrases	Abstracts								RAs							
	Computer science		Linguistics		Biology		Mechanical engineering		Computer science		Linguistics		Biology		Mechanical engineering	
	F	/S	F	/S	F	/S	F	/S	F	/S	F	/S	F	/S	F	/S
NP	566	3.4938	195	3.3051	656	3.8363	513	4.2397	27564	3.2047	10544	3.2255	8504	4.1162	9821	4.0168
PP	139	0.8580	68	1.1525	223	1.3041	196	1.6198	8190	0.9522	3347	1.0239	2705	1.3093	3341	1.3665
PP in NP	73	0.4506	32	0.5424	108	0.6316	104	0.8595	3722	0.4327	1525	0.4665	1206	0.5837	1206	0.6438
VP	242	1.5419	88	1.4938	263	1.4915	198	1.5380	12357	1.4367	4799	1.4680	3390	1.6409	4011	1.6405
/S = per sentence																
Σ sentences	8601		3269		2066		2445		162		59		171		121	

Abstracts: $\chi^2 = 23.6283$, df = 9, p-value = 0.004929
RAs: $\chi^2 = 120.568$, df = 9, p-value < 2.2e-16

Table 5.22: Frequency of occurrence of phrases across disciplines in the AbstRA corpus per sentence

Feature	Abstracts	Research articles
Sentence length	+	
Type/token ratio	+	
Lexical words (focus on nouns)	+	
Lexical density	(-)	
Most frequent lexical items	(+; qualitative)	
Keywords	(+; qualitative)	
Modals		+
Passives	+	
Nominalizations	+	
Grammatical complexity	+	

+ = statistically significant difference
(-) = not significant; statistical test at the border of significance
(+; qualitative) = difference; qualitative analysis

Table 5.23: Summary of the deductive empirical analysis across text type

5.1.4 Summary of the deductive empirical analysis

The results of the quantitative empirical analysis of the ten chosen features comprising shallow, lexical, and grammatical features, can be summarized as displayed in Tables 5.23, across text types, and 5.24, across disciplines. All features were statistically tested for significance, with the exception of two features. The most frequent lexical items and keywords, which although analyzed quantitatively, were not suitable for a statistical test (cf. Sections 5.1.2.2 and 5.1.2.3 for detailed discussion).

According to Table 5.23, abstracts show higher frequencies of all features in comparison to their RAs apart from modals. RAs use modals significantly more frequently than abstracts. Furthermore, all results are statistically significant apart from lexical density. The results for lexical density are formally not significant since the p-value is higher than 0.05 (p-value = 0.0601, cf. Section 5.1.2.1). However, this result shows that there is a strong tendency that abstracts have a higher lexical density than their RAs. This is due to the fact that the p-value still indicates a 93.99% of probability that the difference of the values for lexical density between abstracts and RAs is not due to chance.

Three conclusions can be drawn from these results. The first conclusion is that abstracts are very distinctive types of texts in comparison to their research articles. The results up to this point indicate that abstracts show significantly higher frequencies of occurrences of features that are very typ-

Feature	Variation across disciplines
Sentence length	+
Type/token ratio	+
Lexical words	+
Lexical density	+
Most frequent lexical items	(+; qualitative)
Keywords	(+; qualitative)
Modals	+
Passives	+
Nominalizations	+
Grammatical complexity	+

+ = statistically significant difference
(+; qualitative) = difference; qualitative analysis

Table 5.24: Summary of the deductive empirical analysis across disciplines

ical indicators of *expository* texts. In contrast, the lower frequency of these same features in RA can be interpreted as indicative of properties typical of *argumentative* texts (Biber 1988). Thus, according to the results, abstracts are likely to be *expository* texts and RAs *argumentative* texts.

The second conclusion is that these results can be linked to statistically significant differences in field, tenor and mode of discourse of abstracts and RAs. As shown in Table 4.3 (p. 67), lexical words, modals, nominalizations, passives, and type/token ratio are linguistic features, which are indicative of the sub-category *goal orientation* of the parameter of the context of situation, *field of discourse*. Since all these indicators showed statistically significant differences between abstracts and RAs, it can be interpreted that there is a significant difference in the configuration of the *goal orientation* between these two text types. The same is valid for the sub-categories of parameter *tenor of discourse*, *social role relationship* and *social distance*, which indicators are modals, nominalizations, sentence length, lexical words, type/token ratio, and grammatical complexity, that showed statistically significant differences between abstracts and RAs. However, for the last parameter of context of situation, *mode of discourse* and its two sub-categories, *language role* and *medium*, there is not such a clear distinction between abstracts and RAs as the two former parameters. The main reason for that are the significance results for the indicator lexical density, which are not significant, yet almost at the border of significancy. Hence, it can be concluded that there are not only statistically significant differences between abstracts and RAs, but also that there are main differences in the configurations of the *goal orientation*, as a sub-category of the *field of discourse*, and of the *tenor of discourse*.

The third conclusion drawn from the results of the quantitative empirical analysis so far is that there is definitively a difference across domains within abstracts and within RAs. Generally, biology and mechanical engineering tend to present similar results, while computer science and linguistics tend to differ from the others disciplines quite clearly. These results are indicative of statistically significant differences in the configuration of the parameter of context of situation *experiential domain*, a sub-category of the *field of discourse*. Variation in the *experiential domain* is normally associated with *register* variation (Halliday & Martin 1993), so that it can be said that these four disciplines constitute different registers.

Moreover, both abstracts and RAs are full texts, in the sense that each of them is a complete text, and that they are independent from each other. They make sense by themselves thoroughly and each of them could potentially stand alone. As discussed in Section 4.4, the definition of *goal orientation* is *very similar* to the definition of *genre* according to Martin (1992a). For this reason, it is plausible to infer that differences in the parameter of *goal orientation* most likely reflect *genre variation* (p. 64). Since the quantitative results indicate that abstracts are significantly more *expository* and that RAs are significantly more *argumentative* texts, it is plausible to presume that these two text types are, as a matter of fact, two different *genres variations* (cf. Chapter 6).

In order to verify these interpretations of the results, an inductive empirical analysis, also called bottom-up analysis, is performed with the same data. The two approaches used in this research, hierarchical agglomerative cluster analysis and principal component analysis, are theory independent approaches for analyzing quantitative data requiring no formulation of hypotheses prior to the quantitative analysis. The next section, Section 5.2, discusses these two approaches and presents their results.

5.2 Inductive empirical analysis

Inductive analysis does not require the formulation of hypotheses prior to the quantitative experiment itself. On the contrary, the obtained data are the basis for inferences taken about a given topic. Such analysis allows an investigation of whether the features chosen for analysis are suitable for distinguishing and grouping several samples, in this case, texts. Moreover, the data provided by such an analysis can be used to support the evaluation of how well the chosen features are related to the indicators (cf. Section 4.4).

The inductive approaches applied here, i.e., the hierarchical agglomerative cluster analysis (cf. Section 5.2.1) and for the principal component analysis (cf. Section 5.2.2) are *unsupervised* ones, “in the sense that we do not prescribe what groupings should be there” (Baayen 2008: 118). This analysis is carried out completely in R. The data used for the inductive analysis is shown in Table A.3 in Appendix A.5 (p. 207). Table A.3 is equal to the matrix loaded in R for both approaches used in the hierarchical agglomerative cluster analysis and in the principal component analysis. The script used in R is described in Appendix A.6 (p. 214) . The features taken into consideration in the inductive analysis are:

- anonymized unique name for each single text (row names)
- type
  (abstracts, RAs)
- domain
  (A: computer science, C1: linguistics, C2: biology, C3: mechanical engineering)
- tokens
- prepositions
- adjectives
- modals
- nouns
- personal pronouns
- possessive pronouns
- adverbs
- present tense
  (VB + VBP + VBZ + VHP + VHZ + VVP + VVZ)

- past participle
  (VVN; feature related to passive occurrence; see below)
- past tense
  (VBD + VHD + VVD)
- nominalizations
- sentence length
- lexical density

The features considered in the inductive empirical analysis differ from those chosen for the deductive analysis. The inductive empirical methods used here allow the automatic processing and evaluation of several features at once. Thus, it is now possible to investigate more features than those used in the deductive approach. On the other hand, they require many features, each of them being quantified per sample, i.e., per text, to be analyzed simultaneously, so that their processing algorithm can function properly (Baayen 2008; Crawley 2007). Some of the features chosen here are the same used in the deductive analysis: adjectives, nouns, adverbs, modals, nominalizations, sentence length, and lexical density. Moreover, the total number of tokens per text used here was also indirectly used in the deductive analysis for normalization purposes of the results so that they could be compared. Since the information concerning passives is not available for each single text (cf. Section 5.1.3.2), the values of frequency of occurrence of past participle forms is adopted here. Although not every single occurrence of a past particle implies a passive, each passive occurrence implies obligatorily the existence of a passive verb form, i.e., a past participle. For this reason, it is considered a valid approximation to use the frequency of occurrence of past participles as a correlation to the use of passive voice in the texts under study. In addition to the features considered in the deductive analysis, prepositions, personal and possessive pronouns, present and past tense are taken into consideration here. The use of prepositions is related to the size of nominal and verbal groups being therefore related to grammatical complexity. Finally, as discussed in Section 4.4, verb tenses – present and past tense – are a relevant criteria for characterizing register variation. The values for all features per text used in the inductive analysis can be found in Table A.3. The values in each row of this table are either the raw value of frequency of occurrence of a given feature or a sum of several parts-of-speech indicating a given feature, as for example in the case of present and past tense, or the result of a formula, such as lexical density. As already mentioned in Section 5.1.1.2, there is a text is the abstract corpus, *abstract.C2.5* (and therefore also its correspondent RA, *RA.C2.5*) which is not considered at all for being misbuilt.

5.2.1 Hierarchical agglomerative cluster analysis

Hierarchical agglomerative cluster analysis is an inductive approach comprising several techniques for clustering data (Baayen 2008; Crawley 2007). It aims to group samples based on their similarity / dissimilarity in a space of m-dimensions, where each variable (column) defines a dimension. Similarity is defined "on the basis of the distance between two samples in the m-dimensional space" (Crawley 2007: 742). The default is to calculate the Euclidean distances, i.e., usual square distance between the two vectors, from sample to sample. As described in Appendix A.6, the distance matrix is generated from the scaled and centered initial data matrix. The cluster function in R initially assigns each sample a single cluster. Then, it proceeds iteratively grouping similar clusters together, until there is just one single cluster. The resulting plot is called a dendrogram. The resulting dendrogram for the text samples of the AbstRA corpus is displayed in Figure 5.18. Every single text is plotted as its own cluster at the bottom of the dendrogram in Figure 5.18. Starting from every text, a vertical line is drawn upwards reflecting the degree of similarity of the texts based on its height. The vertical lines are then grouped together by horizontal lines. The longer the vertical lines, the more distinct the clusters are.

According to Figure 5.18, it can be noticed that two text samples, *abstract.A.2* (second abstract in computer science) and *abstract.C3.5* (fifth abstract in mechanical engineering), are very distinct from all other texts in the AbstRA corpus since they were clustered separately from all other text samples. This observation can be interpreted as an indication that these two texts samples are outliers in comparison to the whole AbstRA corpus. More importantly, it can be observed that the other texts are grouped into two main groups: abstracts at the left side and RAs at the right side of the dendrogram. There are only very few mismatches in these two groups. For instance, within the left group – the group of abstracts – there are only 5 *RA.C2* (RAs from biology), 3 *RA.C3* (RAs from mechanical engineering), and only *RA.C1* (RA from linguistics) are clustered together with the abstracts, meaning that they are somehow similar to abstracts. Similarly, within the right group – the group of RAs – there are 1 *abstract.C3* (abstract from mechanical engineering), 1 *abstract.C2* (abstract from biology), 2 *abstract.C1* (abstract from linguistics), and 2 *abstract.A* (abstracts from computer science) were misclustered. Furthermore, it can be observed that the disciplines tend to be grouped together both within the abstracts group as well as in the RAs group. This observation is a good indication of domain specific similarities and differences that were identified in the texts in the AbstRA corpus.

Another way of looking at data as a whole is to plot the distribution of data by every variable against every other. Such plots are called *pairs plots* (Crawley 2007: 740) and they are a good indicative for data grouping and/or separation. Figures 5.19 and 5.20 show the pairs plots for data used for all inductive empirical analyses, i.e., data from the Table A.3 in Appendix A.5 (p. 207), according to text type (Figure 5.19[47]) and domain (Figure 5.20[48]). Both Figures support the former interpretation of the cluster dendrogram (Figure 5.18) showing a grouping of data, both according to text type and domain.

In order to gain more insight into the different groups resulting from the clustering approach, classification trees were generated. Classification trees are good devices for quantitatively discriminating samples according to certain classes. Classification trees are generated through a process called *recursive partitioning*, by which its algorithm inspects the data dividing them into a series of subsets that do not overlap recursively. A detailed description of recursive partitioning can be found in Baayen (2008: 148-154) and Crawley (2007: 695-700). Here, the texts from the AbstRA corpus were classified according to text type, i.e., abstract or RA, and then according to discipline. Classification trees are fully automatically generated by R, i.e. without any interference from the researcher, based on the same data used for clustering (cf. detailed script in Appendix A.6). The resulting classification trees for the AbstRA corpus are shown in Figures 5.21, according to text type, and 5.22, according to domain.

Figure 5.21 reflects a decision procedure for determining the classification of the texts as abstracts or RA. In this tree, each node is labeled with a rule. The positive answer to the rule is the left branch of the tree. Thus, it can be seen that for the first distinguishing criterium between the samples, i.e., *modals*, 58 abstracts out of initially 93 are grouped on the first left branch and no RA out of the initial 94 is in this group, i.e., 58/0. This means that the frequency of occurrence of modals is a strong discriminator between abstracts and RAs. The next distinctive feature is *possessive pronoun*, by which initially there are 35 abstracts to be classified and 94 RAs

[47] Black = abstracts, red = research articles

[48] Black = computer science, red = linguistics, green = biology, blue = mechanical engineering

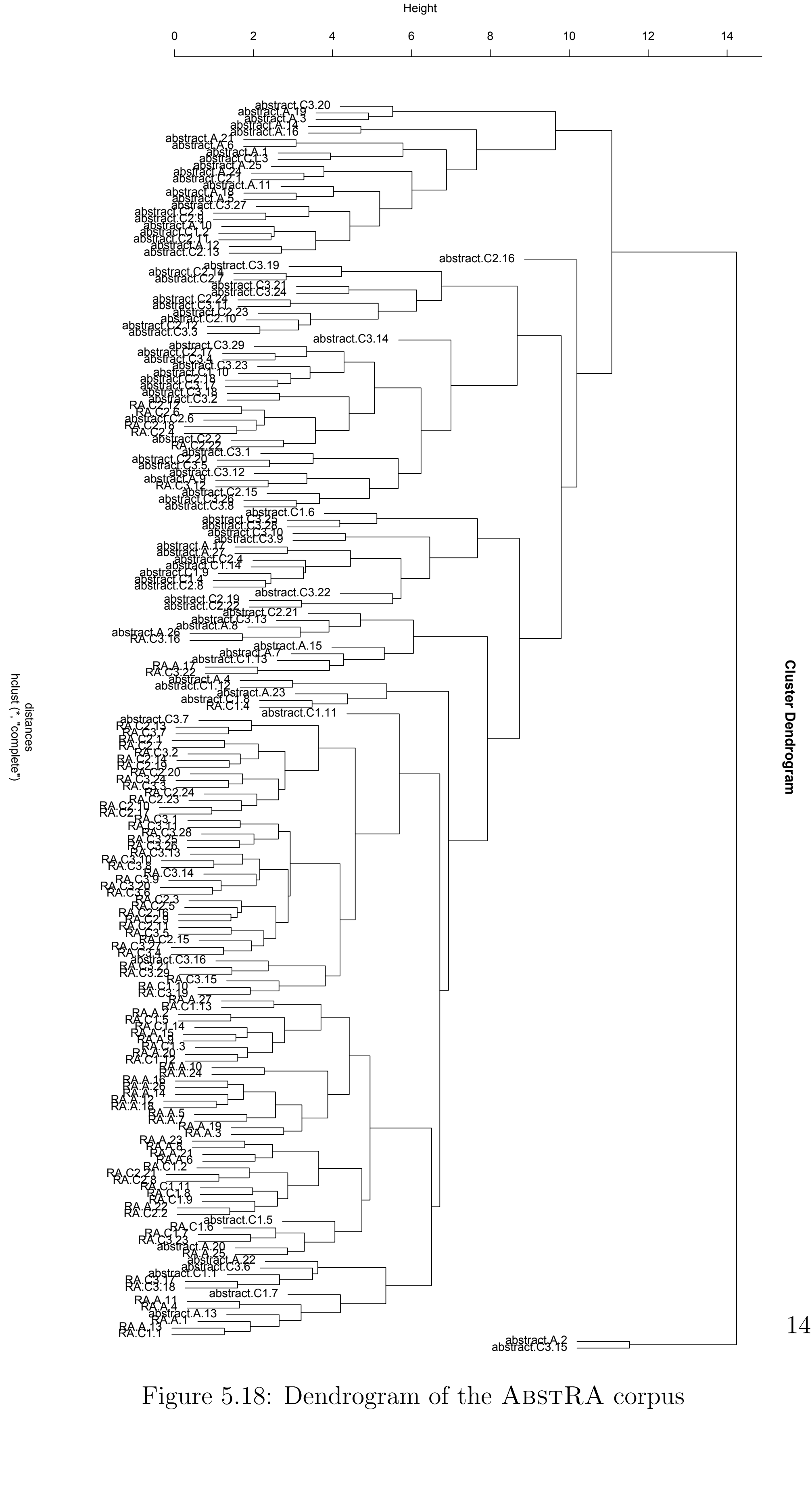


Figure 5.18: Dendrogram of the AbstRA corpus

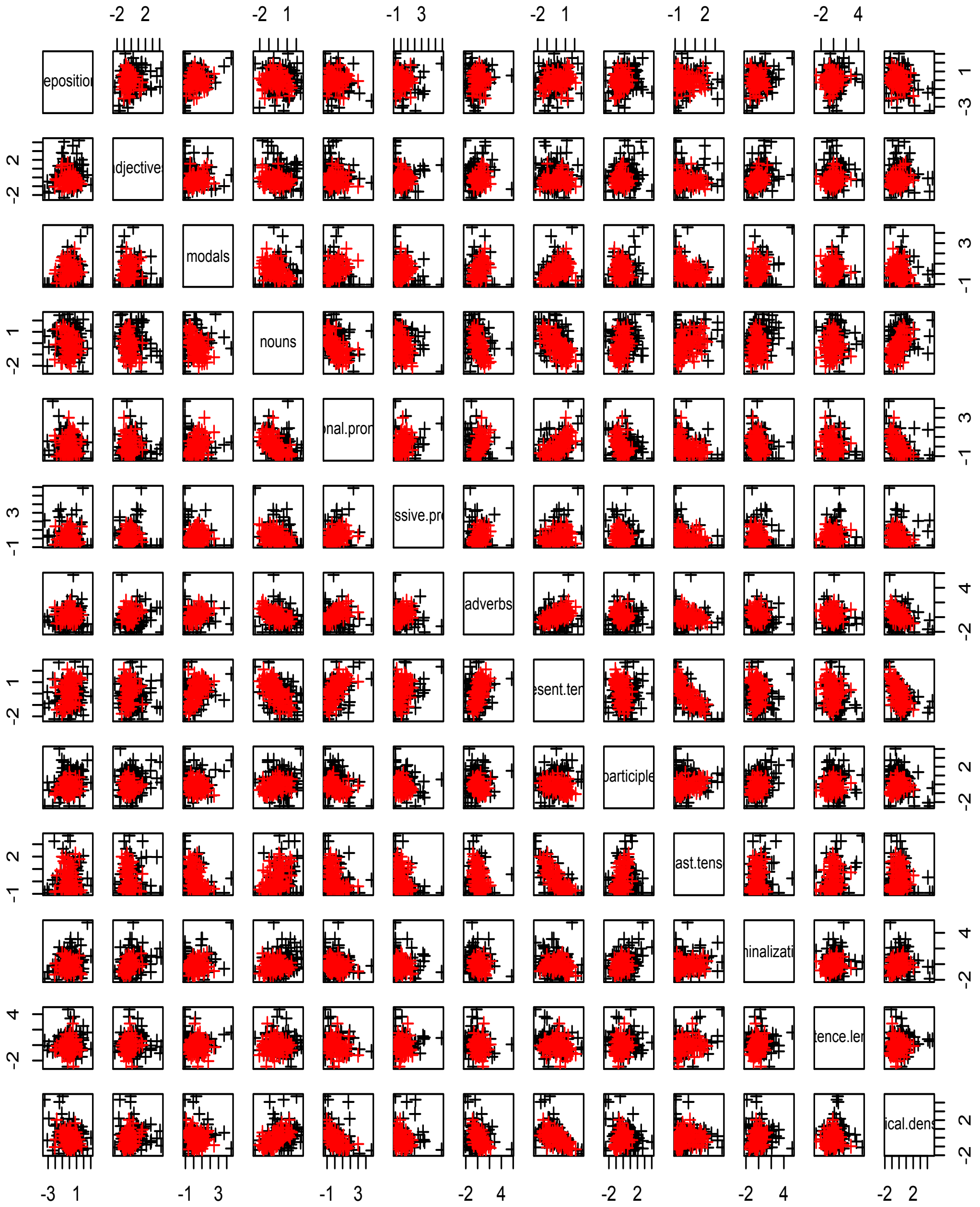


Figure 5.19: Pairs plot of all features analyzed in the texts of the AbstRA corpus against each other according to text type

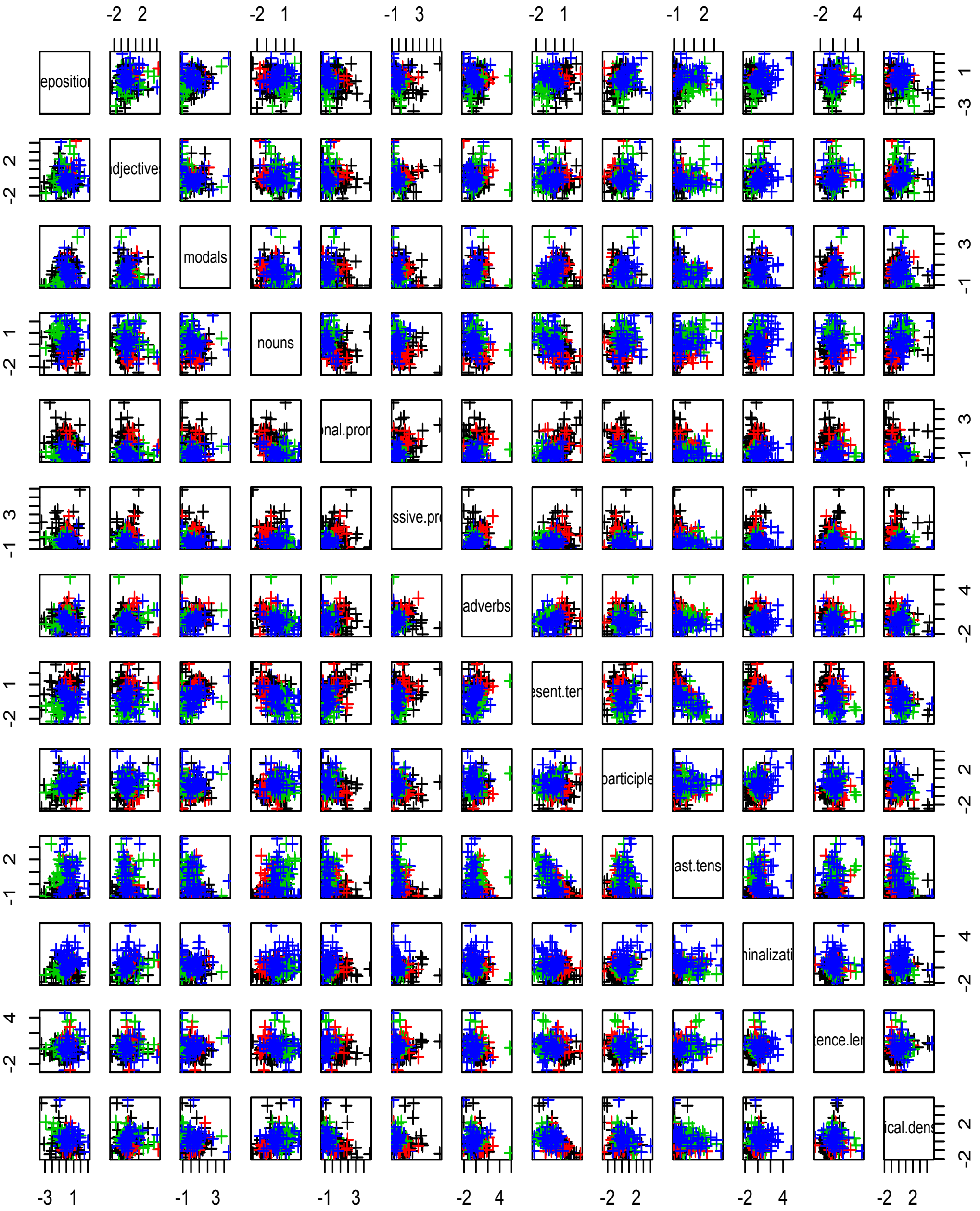


Figure 5.20: Pairs plot of all features analyzed in the texts of the AbstRA corpus against each other according to domain

(35/94). According to Figure 5.21, there are 26 out of the 35 abstracts that are grouped with this criterium. However, 4 out of the 94 RAs are also classified into this group. It should be noticed though, that it is very unusual that all texts would be grouped ideally separated from each other. Furthermore, only with these two criteria, 84 out of 93 abstracts are already distinguished from RAs, i.e., 90.32% of all abstracts. The classification process goes on and the remaining 9 abstracts and 90 RAs are classified according to the feature *past tense*. The left branch of this last node contains 5 abstracts and 2 RAs and the right branch contains 88 RAs (93.62%) and only 4 abstracts. Therefore, the features automatically chosen by the classification tree feature in R – modals, possessive pronouns, and past tense – are very adequate for distinguishing between abstracts and RAs. In other words, abstracts and RAs are very distinctive text types. Moreover, according to the classification tree in Figure 5.21, the features *modals*, *possessive pronouns*, and *past tense* are the most distinctive ones for discriminating between these two different text types: abstracts and RAs.

The same rationale can be applied to the second classification tree in Figure 5.22, which resulted from the classification of the texts in the AbStRA corpus according to domain. According to Figure 5.22, there are initially 54/58/47/58 texts; abstracts *and* RAs, i.e., 54 texts from domain A (computer science); 28 texts from domain C1 (linguistics); 47 texts from the domain C2 (biology); and 58 texts from domain C3 (mechanical engineering). The first node of the tree classifies the texts based on the feature *personal pronouns*. The left branch of this tree groups 42/21/10/7 texts, i.e., 77.77% of the texts from computer science, 75.00% of the texts from linguistics, 21.28% of the texts from biology, and 12.07% of the texts from mechanical engineering. This already indicates that texts from computer science and linguistics are very distinct from texts from biology and mechanical engineering. Besides, computer science and linguistics are also very distinct since A is on the left branch and C1 is on the right branch of this first node of the tree. Furthermore, on this left branch of this first node, the remaining texts are additionally classified according to *nouns*. Here, the few misclassified texts are again separated into two distinct groups: C1 (linguistics; 2/10/1/2) and the very few of other domains, mainly C2 (biology; 3/1/6/4). The right side of the branch of the first node of Figure 5.22 can be analyzed similarly. The feature *nominalization* separates two main groups from the initial 12/7/37/51 texts: 12 texts from computer science (22.22%); 7 texts from linguistics (25.00%); 37 texts from biology (78.72%); and 51 (87.93%) texts from mechanical engineering. The feature nominalization separates 62.71% of the texts of C2 (29 texts of biology) from 63.79% of the

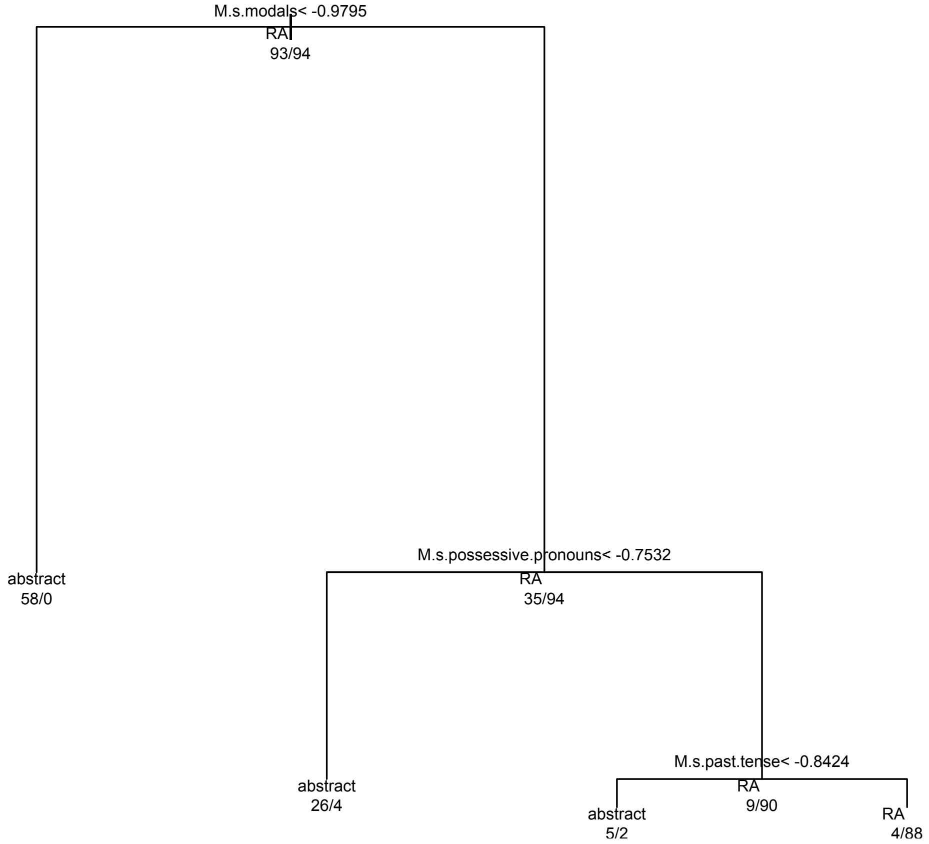


Figure 5.21: Classification tree of the AbstRA corpus according to text type

texts of C3 (37 texts of mechanical engineering). Within C2 there are some misclassified texts from A, i.e., computer science. The left branch of this node shows a distinct separation according to *past tense*. This feature distinguishes then the remaining texts from computer science (A) very clearly from biology (C2). Similarly, the few texts from linguistics (C1) and the 34 texts from C3 (mechanical engineering) are grouped separately according to the feature *past participle*. Thus, there is a clear domain specific grouping of the texts from the AbstRA corpus for each single discipline. Moreover, while computer science and linguistics seem to be more similar to each other, biology and mechanical engineering show more similarities.

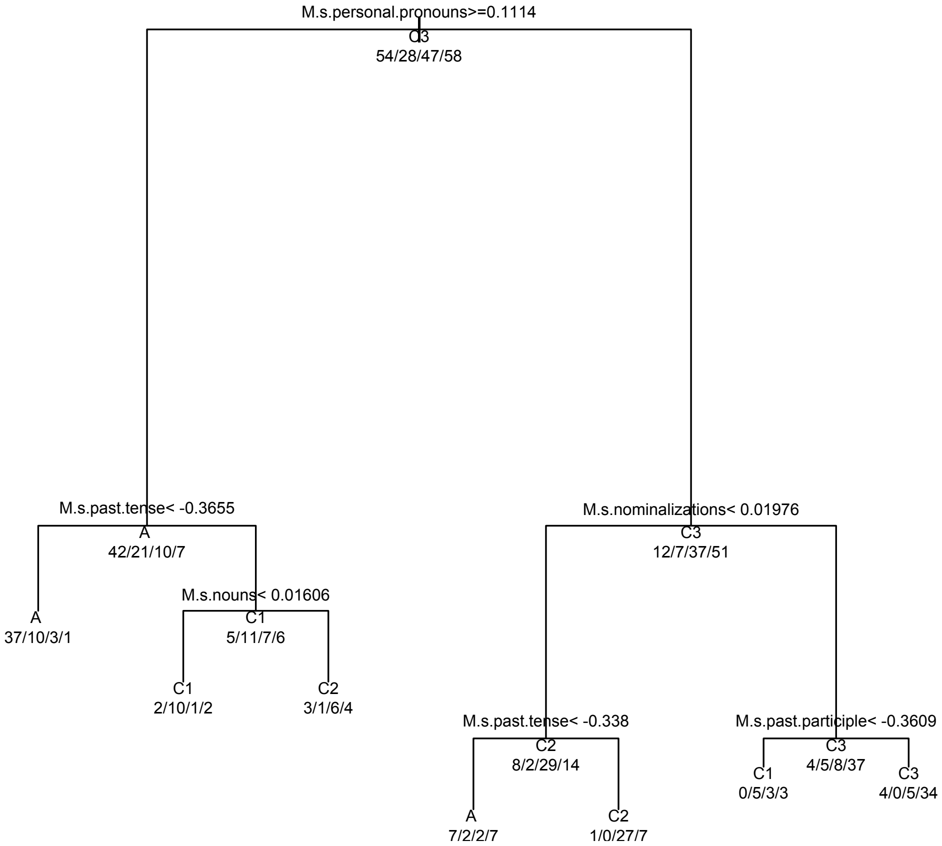


Figure 5.22: Classification tree of the AbstRA corpus according to domain

In order to investigate how the features chosen for the inductive analysis, i.e., the columns of the Table A.3 in Appendix A.5 relate to each other, the data was clustered with the transposed matrix. The resulting dendrogram is found in Figure 5.23. This figure shows two very distinctive clusters. The left one comprises two clusters: one with possessive pronouns, personal pronouns, and present tense, and the other with modals and adverbs. The other main cluster is also composed of two subclusters. The first one comprises the features of adjectives, prepositions, past participle and nominalizations. Finally, the last cluster comprises the features nouns, lexical density, past tense, and sentence length.

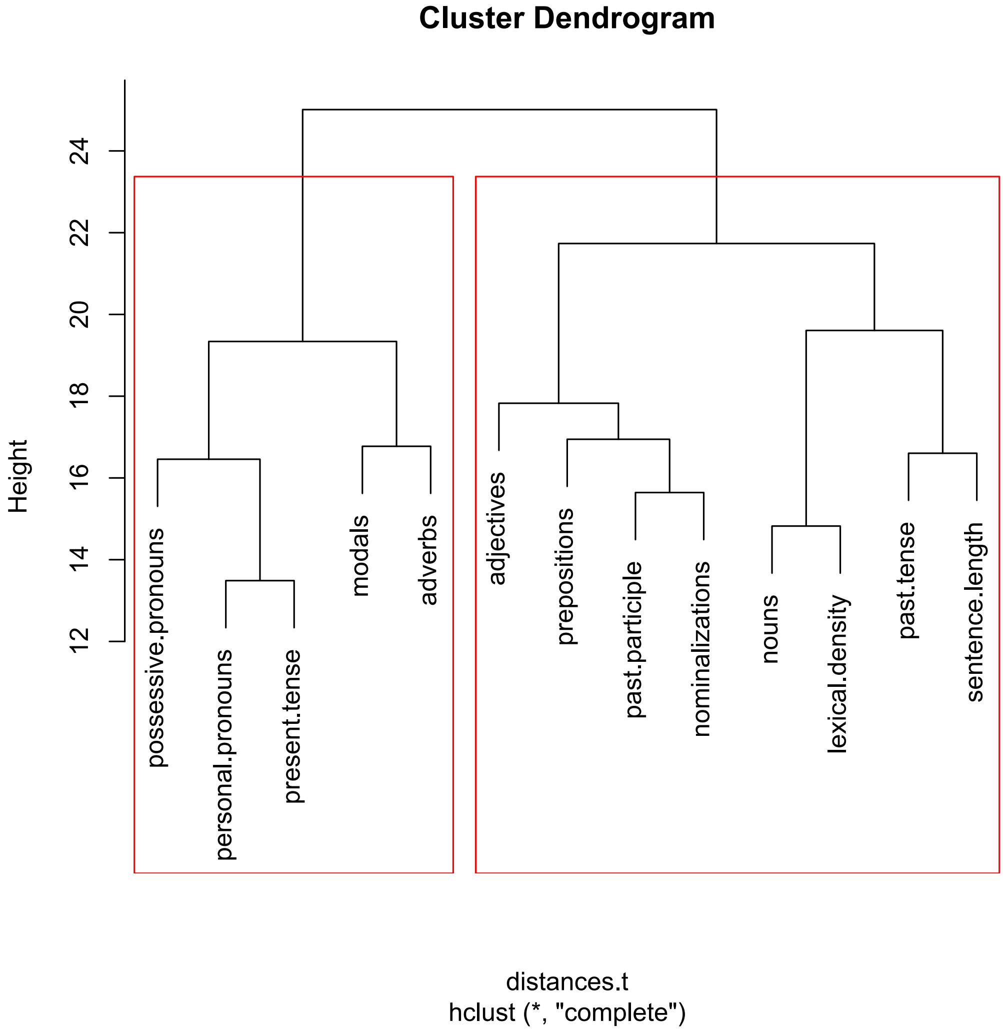


Figure 5.23: Dendrogram of the features analyzed in the texts of the AbstRA corpus

The striking question about this clustering is to find out why these features are clustered together and how they relate to each other. When comparing the results of Figure 5.23 with the results found by the classification tree, it can be noticed that the left main cluster also includes the most distinctive features between text types (cf. Figure 5.21). Similarly, the right cluster of Figure 5.23 comprises the features for major distinction of the texts in the AbstRA corpus according to domain (cf. Figure 5.22). The next challenge is to interpret this data and infer how such feature clustering into two main groups, i.e., features for text type clustering and features for domain clustering, relate to the original indicators of the contextual parameters field, tenor and mode of discourse, which are the theoretical background for characterizing language and language variation adopted in this research (cf. Section 4.4). In order to investigate this issue, a last inductive approach is presented in this research: principal component analysis, which is discussed in details in the next section.

5.2.2 Principal component analysis

Principal component analysis (PCA) is a technique for analyzing multidimensional data. It aims to reduce the number of dimensions of data finding common dimensions in the set of data variables. PCA allows the researcher to find patterns in data and to compress data by such a reduction of the number of dimensions, helping the process of data interpretation. Such reduced space dimensions may be interpreted as categories. This is because these new principal components can be interpreted as clusters themselves. Similarly, the original values of data which were reduced and projected in few new dimensions can be seen as a reflection of the membership degree of the original values in each cluster (Baayen 2008; Crawley 2007). Principal component analysis in R is very straightforward (cf. detailed script in Appendix A.6). The interpretation of the numerical data given by the principal components is however rather laborious.

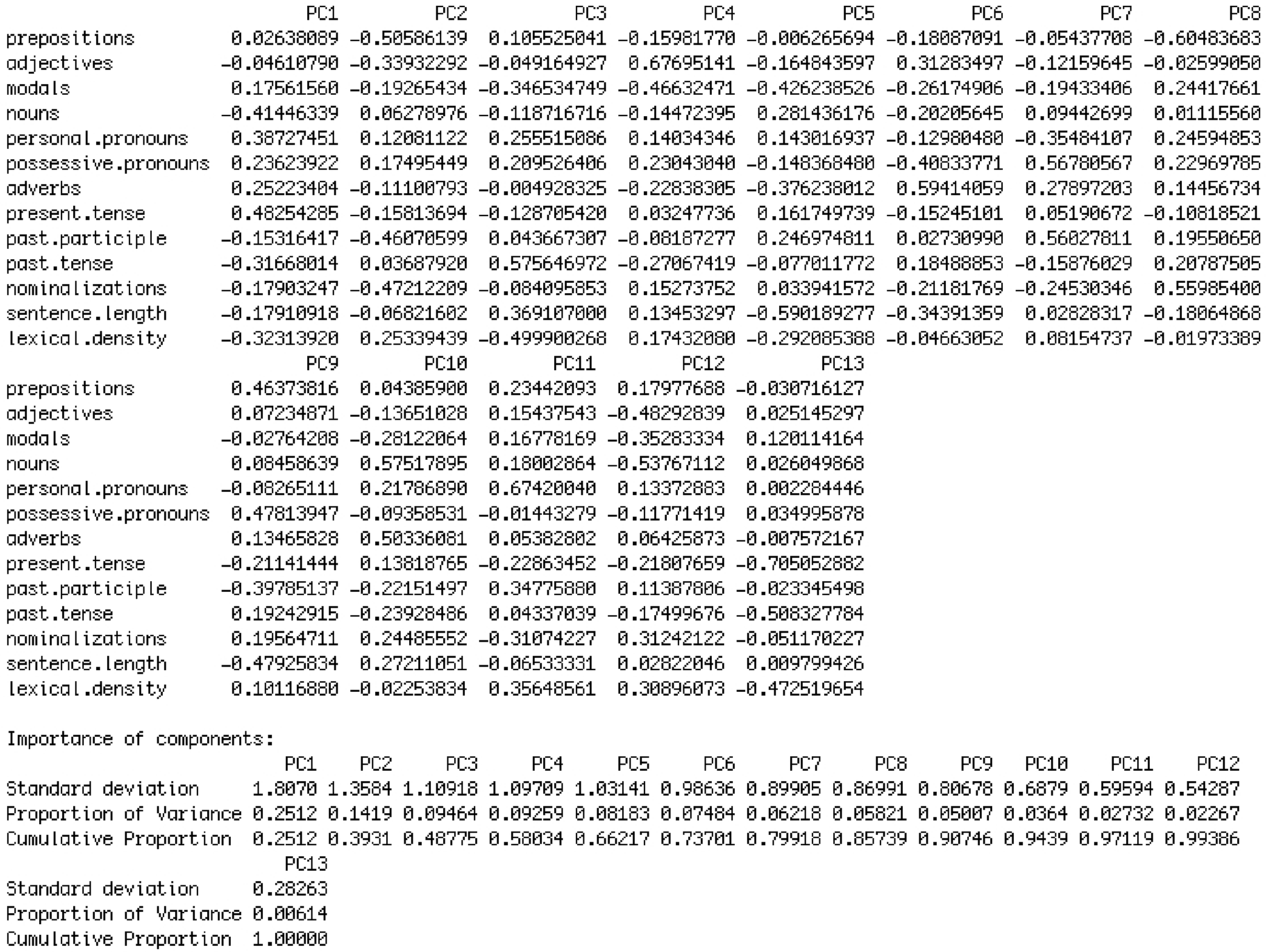
The result of the PCA for the data in Table A.3 (p. 213), i.e., the same data used as input for the hierarchical agglomerative cluster analysis, is displayed in Figure 5.24. According to these results, the first principal component (PC1) explains 25.1% of the total variation, the second principal component (PC2) explains 14.2% of the variance, the third principal component (PC3) explains 9.46% of the variance, and so on. The standard procedure is to consider as many PCs needed to account for 90% of the total variation (Crawley 2007: 733). However, this would take 9 components that would represent 9 dimensions. The distribution of the total variance across the principal components can be seen in the so called *scree plot*[49] in PCA, as shown in Figure 5.25. This curve looks like a cliff with slope below it. According to this figure, it can be seen that PC3 and PC4 explain almost in equally percentage the total variance and that the slope in general is not very abrupt from PC3 on. There is a rule of thumb to locate the cutoff point on how many principal components to consider, which says that this is the point where a clear discontinuity is seen, form the right to the left of the Figure 5.25 (cf. Baayen 2008: 121). According to this rule, the cutoff point would be after PC3. Hence, only PC1, PC2, and PC3 are to be considered in the further analysis.

The next step in the analysis of these results is an interpretation of the numerical data concerning PC1, PC2, and PC3 in Figure 5.24. The first principal component, PC1, has high positive loadings of modals, personal pronouns, possessive pronouns, adverbs, and present tense. This data match precisely the left main cluster obtained by the hierarchical agglomerative cluster analysis (cf. Figure 5.23). PC1 also shows high negative loadings of nouns, past tense, lexical density, sentence length, and nominalizations. These negative loadings match with the right main cluster obtained by the hierarchical agglomerative cluster analysis (cf. Figure 5.23). The features of PC1 with positive loadings: modals, personal pronouns, possessive pronouns, adverbs, and present tense, are very typical for characterizing argumentative discourse. Meanwhile, its features with negative loadings, i.e., nouns, past tense, lexical density, sentence length, and nominalizations, are typical for expository discourse (cf. Section 4.4; similar to the second dimension in the work of Biber (1988), narrative vs. non-narrative concerns).

The second principal component in Figure 5.24, PC2, has lexical density as the main positively loaded feature, and several highly negative loaded features such as prepositions, nominalizations, past participle, and adjectives. These features are very similar to the features in Biber's first dimension, involved vs. informational production, where positive loadings of nouns, prepositions, adjectives, etc., indicate a very informational discourse and careful integration of information in text (Conrad & Biber 2001: 24). The data for PC2 matches also Biber's fifth dimension: abstract vs non-abstract style, where the high negative loadings can be associated with a non-abstract style.

The third principal component in Figure 5.24, PC3, has high positive loadings for the features sentence length, past tense, personal pronouns and

[49] The name is indeed *scree plot* and *not* scree<u>n</u> plot. Detailed information about the *scree plot* is found in Crawley (2007).



```
                            PC1         PC2          PC3         PC4          PC5         PC6         PC7         PC8
prepositions         0.02638089 -0.50586139  0.105525041 -0.15981770 -0.006265694 -0.18087091 -0.05437708 -0.60483683
adjectives          -0.04610790 -0.33932292 -0.049164927  0.67695141 -0.164843597  0.31283497 -0.12159645 -0.02599050
modals               0.17561560 -0.19265434 -0.346534749 -0.46632471 -0.426238526 -0.26174906 -0.19433406  0.24417661
nouns               -0.41446339  0.06278976 -0.118716716 -0.14472395  0.281436176 -0.20205645  0.09442699  0.01115560
personal.pronouns    0.38727451  0.12081122  0.255515086  0.14034346  0.143016937 -0.12980480 -0.35484107  0.24594853
possessive.pronouns  0.23623922  0.17495449  0.209526406  0.23043040 -0.148368480 -0.40833771  0.56780567  0.22969785
adverbs              0.25223404 -0.11100793 -0.004928325 -0.22838305 -0.376238012  0.59414059  0.27897203  0.14456734
present.tense        0.48254285 -0.15813694 -0.128705420  0.03247736  0.161749739 -0.15245101  0.05190672 -0.10818521
past.participle     -0.15316417 -0.46070599  0.043667307 -0.08187277  0.246974811  0.02730990  0.56027811  0.19550650
past.tense          -0.31668014  0.03687920  0.575646972 -0.27067419 -0.077011772  0.18488853 -0.15876029  0.20787505
nominalizations     -0.17903247 -0.47212209 -0.084095853  0.15273752  0.033941572 -0.21181769 -0.24530346  0.55985400
sentence.length     -0.17910918 -0.06821602  0.369107000  0.13453297 -0.590189277 -0.34391359  0.02828317 -0.18064868
lexical.density     -0.32313920  0.25339439 -0.499900268  0.17432080 -0.292085388 -0.04663052  0.08154737 -0.01973389
                            PC9        PC10        PC11        PC12         PC13
prepositions         0.46373816  0.04385900  0.23442093  0.17977688 -0.030716127
adjectives           0.07234871 -0.13651028  0.15437543 -0.48292839  0.025145297
modals              -0.02764208 -0.28122064  0.16778169 -0.35283334  0.120114164
nouns                0.08458639  0.57517895  0.18002864 -0.53767112  0.026049868
personal.pronouns   -0.08265111  0.21786890  0.67420040  0.13372883  0.002284446
possessive.pronouns  0.47813947 -0.09358531 -0.01443279 -0.11771419  0.034995878
adverbs              0.13465828  0.50336081  0.05382802  0.06425873 -0.007572167
present.tense       -0.21141444  0.13818765 -0.22863452 -0.21807659 -0.705052882
past.participle     -0.39785137 -0.22151497  0.34775880  0.11387806 -0.023345498
past.tense           0.19242915 -0.23928486  0.04337039 -0.17499676 -0.508327784
nominalizations      0.19564711  0.24485552 -0.31074227  0.31242122 -0.051170227
sentence.length     -0.47925834  0.27211051 -0.06533331  0.02822046  0.009799426
lexical.density      0.10116880 -0.02253834  0.35648561  0.30896073 -0.472519654

Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9   PC10    PC11    PC12
Standard deviation     1.8070 1.3584 1.10918 1.09709 1.03141 0.98636 0.89905 0.86991 0.80678 0.6879 0.59594 0.54287
Proportion of Variance 0.2512 0.1419 0.09464 0.09259 0.08183 0.07484 0.06218 0.05821 0.05007 0.0364 0.02732 0.02267
Cumulative Proportion  0.2512 0.3931 0.48775 0.58034 0.66217 0.73701 0.79918 0.85739 0.90746 0.9439 0.97119 0.99386
                          PC13
Standard deviation     0.28263
Proportion of Variance 0.00614
Cumulative Proportion  1.00000
```

Figure 5.24: Principal component analysis of the features analyzed in the texts of the AbstRA corpus

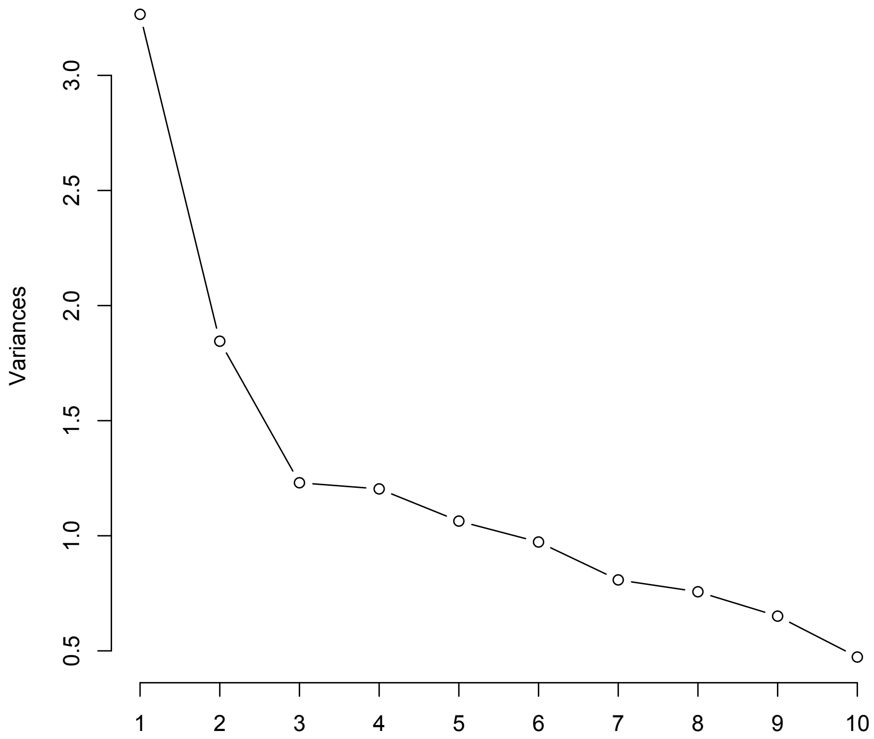


Figure 5.25: *Scree plot* in PCA

possessive pronouns and high negative loadings for lexical density, modals, nouns, and present tense. These data can be associated with Biber's fourth dimension, overt expression of persuasion/argumentation, by which negative loadings of these features would represent not overtly argumentative texts (Conrad & Biber 2001: 36).

In order to better visualize how features and principal components relate to each other, a *variables factor map* is generated, as shown in Figure 5.26. This plot is however only bi-dimensional; therefore only the first two principal components are plotted against each other. As presented in Figure 5.26, there is a clear distinction between two groups of features, which are identical to the two main clusters found in Figure 5.23. This figure corroborates the previous interpretation of the numerical data.

However, principal component analysis is really attractive for the insights offered when the dimensions and components are plotted in several other ways (cf. Crawley 2007). First of all, scatterplot matrices plot the

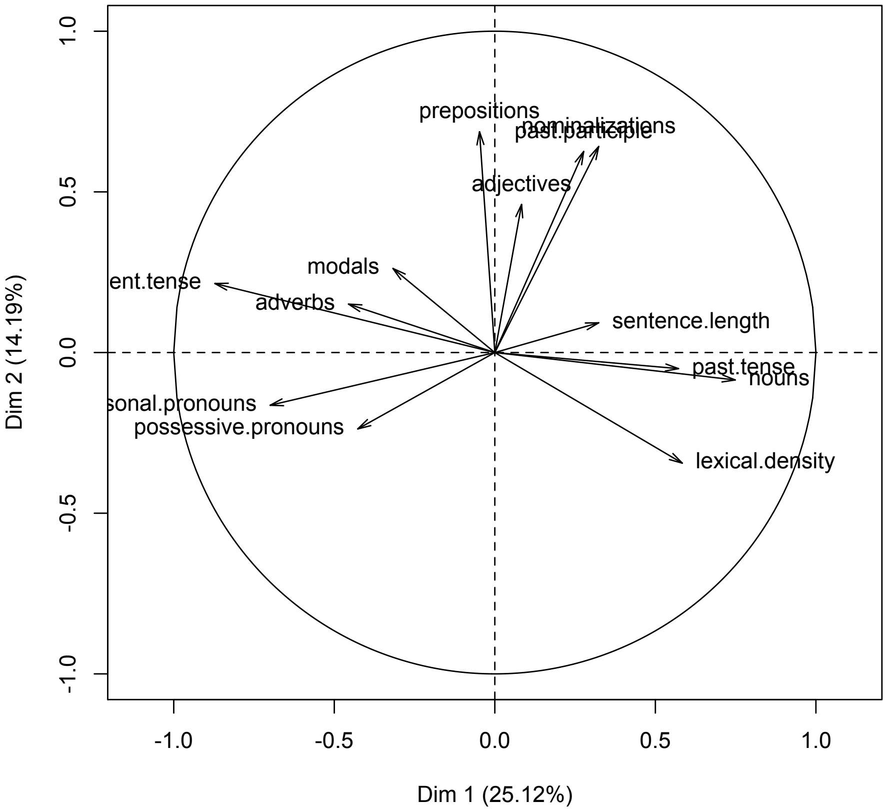


Figure 5.26: Variables factor map for the two first principal components of the PCA

distribution of texts in the multidimensional space created by the principal components. An example of such an scatterplot matrix is found in Figure 5.27, where the texts of the AbstRA corpus are classified according to the three first components. The focus is on the two different text types, abstracts in black and RAs in red. This figure simulates looking at a cube from three different sides, from the top, form the front, and from the other side (Baayen 2008: 122). This figure shows that PC2 offers a very good distinction between abstracts and RAs. PC2 contains high positive loadings for lexical density and highly negative loaded features for prepositions, nominalizations, past participle, and adjectives. As discussed before, these can be associated with distinctions between very informational and abstract discourse and their counterparts. It should be kept in mind, however, that each plot is a compression of data from a multidimensional space into fewer dimensions, always resulting in compression of data. This could be a reason for apparently overlapping red and black dots.

In an attempt to better visualize how the texts are distinguished according to text type, i.e., whether they are abstracts ore RAs, a 3-D scatterplot for the first three principal components is generated as shown in Figure 5.28. This figure indicates that abstracts and RAs can be clearly distinguished from each other since the several dots apparently *do not* overlap. Apparently, abstracts are somewhat *floating around* RAs. In order to check whether this is true or not, a two dimensional plot using another approach, the *multidimensional scaling* (cf. Baayen 2008: 136), is obtained, as shown in Figure 5.29. The multidimensional scaling approach adopted here creates a bi-dimensional representation of the n-dimensional data, i.e., all data in Table A.3, according to their similarities. The resulting plot show clearly that abstracts – black dots – are almost completely positioned *around* the RAs – red dots. Hence, in the three dimensional plot the black dots, i.e., abstracts, are indeed mainly *orbiting* around the red dots, i.e., RAs. Therefore, it can be inferred that abstracts and RAs are two distinct text types, due to the very few overlaps between red and black dots both in 2-D and 3-D visualization of data.

Hence, all results from the inductive empirical analysis, i.e., hierarchical agglomerative cluster analysis, classification trees, and principal component analysis, so far showed that there is a clear distinction between abstracts and RAs as different text types.

The same procedure for the interpretation of the principal component analysis is repeated, but this time concentrating in the differences *across disciplines* and how effective the texts of the different domains are grouped based on such features. The data used is exactly the same set of data used for the former analysis, i.e., the numeric data from Figure 5.24. The only parameter that is changed is that R should now plot the texts showing their *domains* and not whether they are abstracts or RAs. The following colors are used for distinguishing between domains: black for computer science, red for linguistics, green for biology, and blue for mechanical engineering.

The scatterplot matrix for the texts distributed in the multidimensional space spanned by the three first principal components of the PCA across domains is shown in Figure 5.30. It can be first observed that generally, blue & green dots and red & black dots tend to be grouped near each other, specially in PC1 and PC2. This is already an indication that the discourse of mechanical engineering is more similar to the discourse of biology, as well as more dissimilar to linguistics and computer science which are then themselves more similar or nearer to each other. This behavior corroborates the findings of the empirical deductive analysis (cf. Section 5.1) concerning domain variation of data.

Again, PC1 has the following features with positive loadings of modals, personal pronouns, possessive pronouns, adverbs, and present tense. PC1 also show high negative loadings of the features: nouns, past tense, lexical density, sentence length, and nominalizations. As discussed in Section 5.1, the disciplines of biology and mechanical engineering tend to use more nouns and nominalizations and to construe longer sentences with high values of lexical density. Thus, following the interpretation of the principal components PC1, PC2, and PC3 adopted previously, it can be said that these two disciplines – mechanical engineering and biology – show characteristics of a more informational and not very persuasive discourse in comparison to the other two disciplines: computer science and linguistics.

In order to investigate how the texts are distributed in a three dimensional space, i.e., how the relate in space to PC1, PC2 and PC3, a 3-D scatterplot is shown in Figure 5.31. It can be clearly seen that there are few overlapping; hence the four disciplines are very distinct from each other. Furthermore, there is almost a very clear (vertical) cut between the two main groups, green & blue dots and red & black dots. Again, the three dimensional plot shows a higher similarity between texts from biology and mechanical engineering in comparison to computer science and linguistics. Finally, a reduction to a two dimensional plot is performed using the multidimensional scaling approach in order to better visualize the overlapping texts in Figure 5.32. Once more, there are very few overlapping, mainly between red and black dots and some overlapping between blue and green dots. However, this approach allowed the visualization of a very clear distinction between the texts according to their domains. Thus, it can be concluded that texts from different disciplines show distinct linguistic profiles, based on the studied features. Hence, these results corroborate the domain specific register variation expected to be detected in the scientific discourse of different disciplines (cf. Chapter 2).

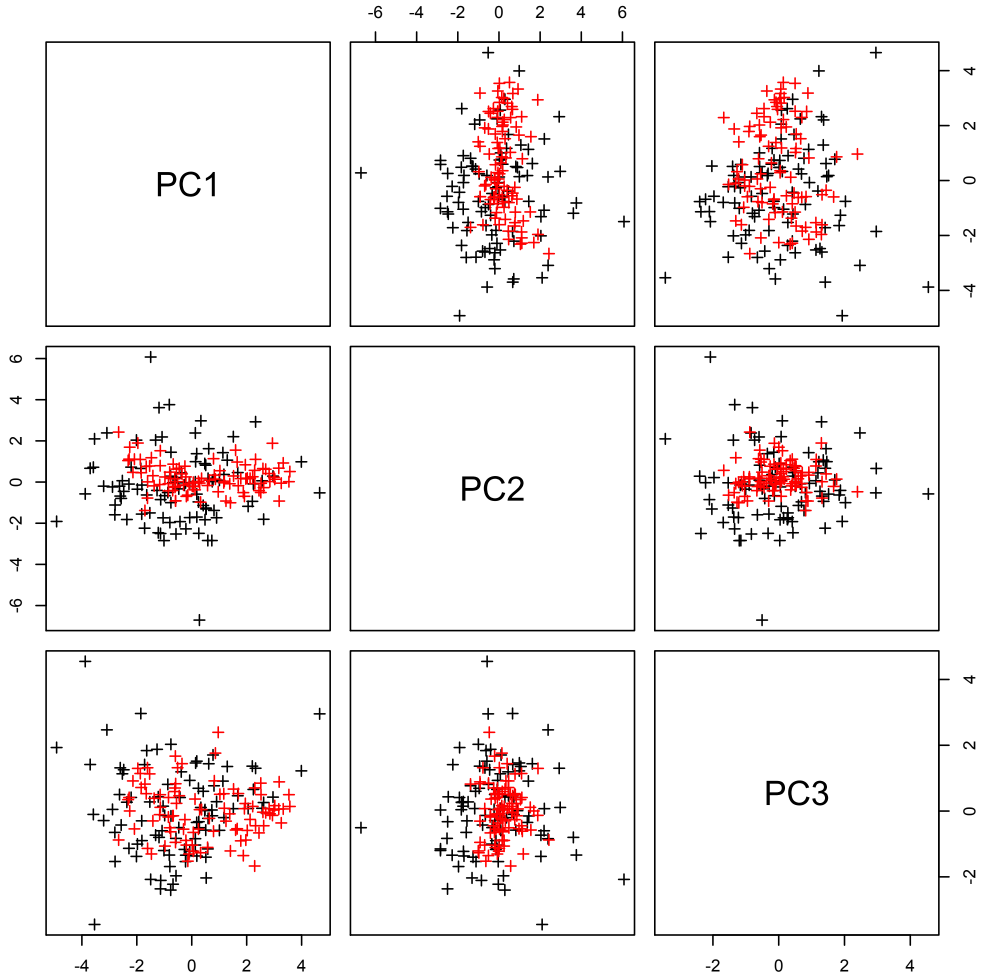


Figure 5.27: Scatterplot matrix for the distribution of texts in the multi-dimensional space spanned by the three first principal components of the PCA according to text type

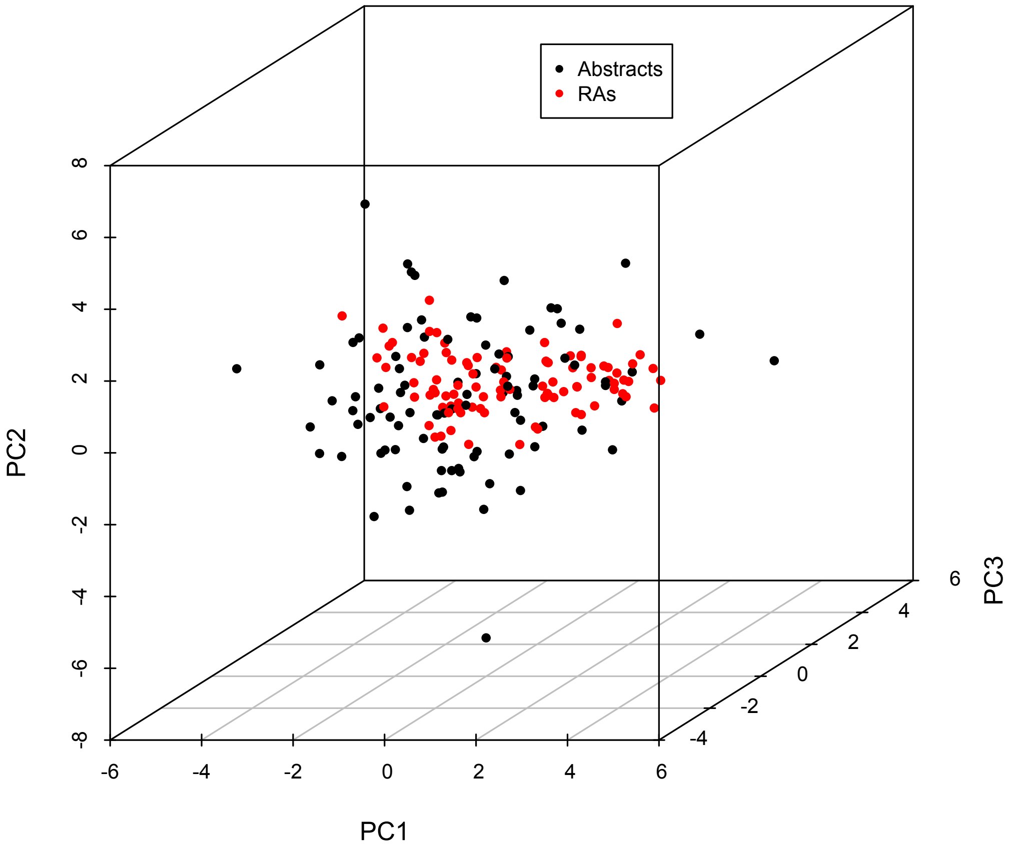


Figure 5.28: 3-D scatterplot for the distribution of texts in the multidimensional space spanned by the three first principal components of the PCA according to text type

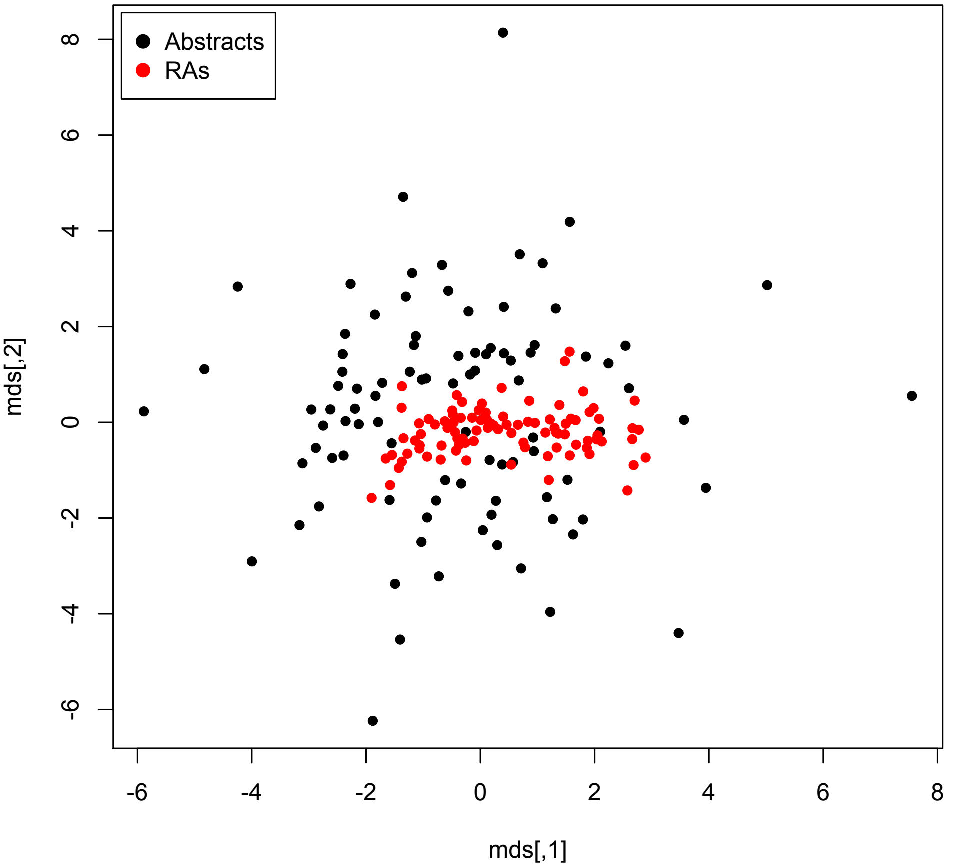


Figure 5.29: Multidimensional scaling of the AbstRA corpus according to text type

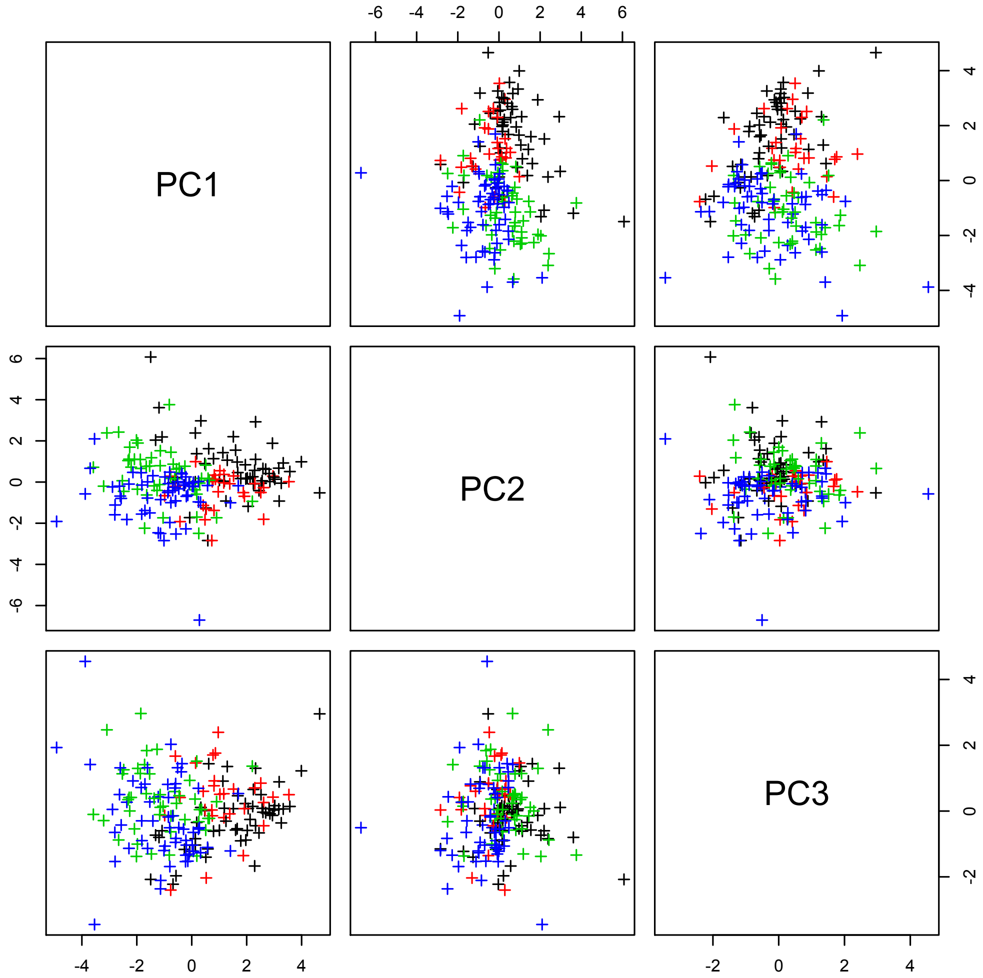


Figure 5.30: Scatterplot matrix for the distribution of texts in the multidimensional space spanned by the three first principal components of the PCA across domains

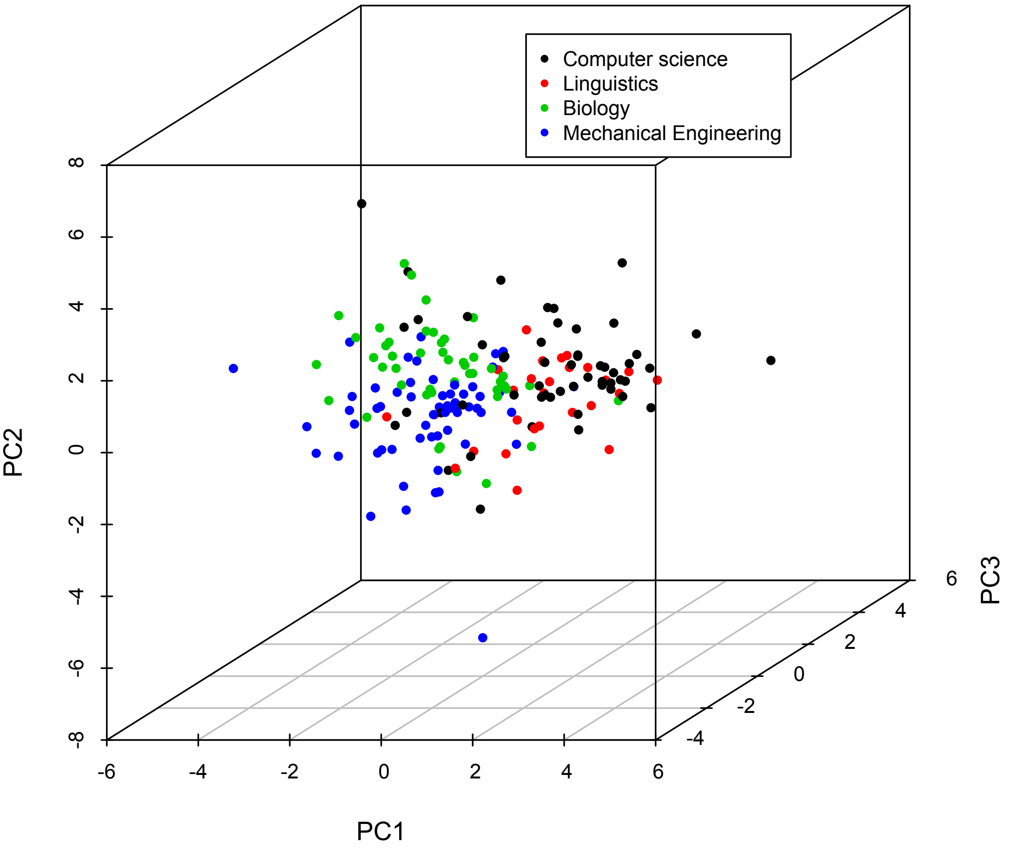


Figure 5.31: 3-D scatterplot for the distribution of texts in the multidimensional space spanned by the three first principal components of the PCA across domains

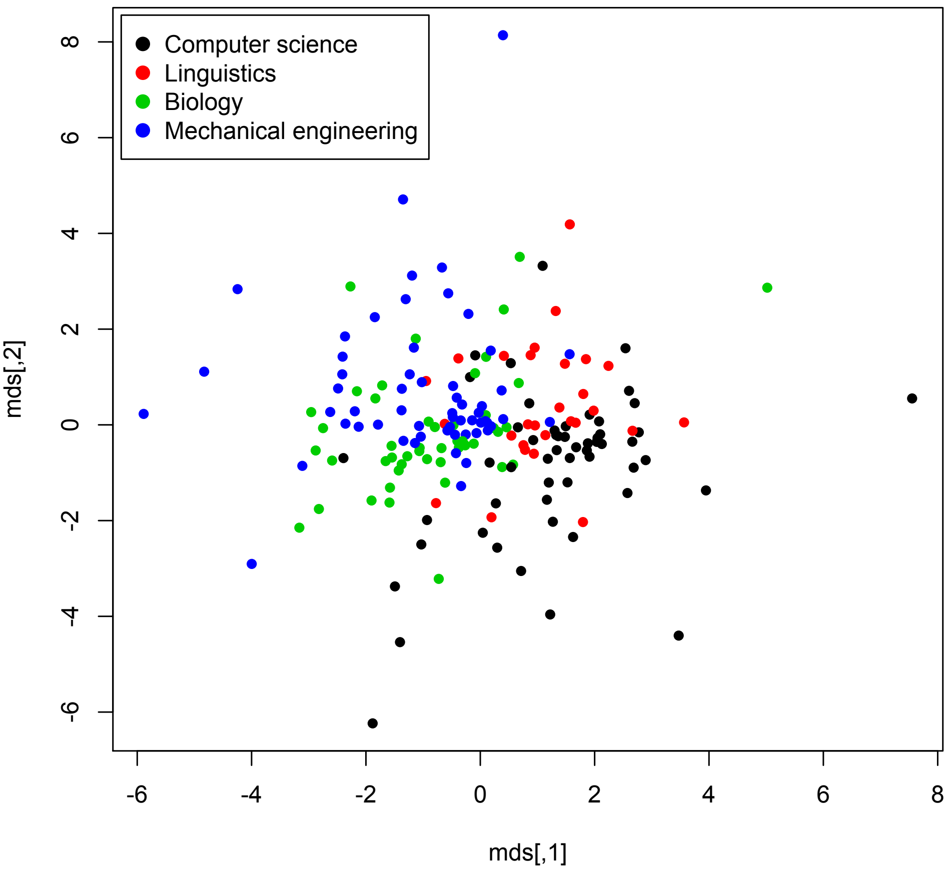


Figure 5.32: Multidimensional scaling of the AbstRA corpus across domains

5.2.3 Summary of the inductive empirical analysis

The methods for inductive empirical analysis adopted in this research, hierarchical agglomerative cluster analysis and principal component analysis, provided insight into similarities and dissimilarities of the texts of the AbstRA corpus, without prior formulation of hypothesis. The approaches used here are based only on the quantitative values of the 13 chosen features: prepositions, adjectives, modals, nouns, personal pronouns, possessive pronouns, adverbs, present tense, past participle, past tense, nominalizations, sentence length, and lexical density.

The results showed that the texts of the AbstRA corpus are clearly grouped according to their text type, i.e., whether they are abstracts or RAs. There are a few text misclassifications in the hierarchical agglomerative cluster analysis and in the classification trees. Thus, it can be concluded that abstracts and RAs are *distinct text types*. This observation corroborates the findings in the deductive empirical analysis as discussed in Section 5.1. Furthermore, there is a distinction of the texts of the AbstRA corpus according to their domains. Similar to the results found in Section 5.1, the inductive methodology applied in this research showed that generally biology and mechanical engineering tend to present similar results. Meanwhile, it is quite evident that computer science and linguistics tend to be separated from the other disciplines: biology and mechanical engineering.

Overall, the obtained data indicate that abstracts are more expository texts, with a high informational and abstract discourse, based on the positive and negative loadings for the features investigated. However, RAs show more properties of argumentative and persuasive texts by which the author is more involved in the text in comparison to abstracts. Such observations can be related to the parameters of field, tenor, and mode of discourse, which were initially proposed as parameters for the investigation of language variation (cf. Table 4.3). For instance, the domain specific variation observed in the set of data (cf. Figure 5.31) can be interpreted as a consequence of variation in the field of discourse, more specifically in the *experiential domain*. This is because the *experiential domain*, as a subparameter of the field of discourse, refers to what is happening, to the topic and to the nature of the action taking place in the discourse (cf. Section 2.4.2). A typical feature that is an indicator for the *experiential domain* is the distribution of nouns, which proved here to be a distinctive feature among the different disciplines (cf. Table 4.3).

On the other hand, the variation according to text type can also be interpreted as correlating to variation in the field of discourse; however more specifically in the *goal orientation*. As discussed in Section 4.4, typical features functioning as indicators for variation in the parameter *goal orientation* are, for instance, lexical words, modals, nominalizations, type/token ratio, pronouns and passives (substituted here for past particle), (cf. Section 5.2; Table 4.3). All these features contributed in the inductive analysis through their positive and negative loadings to a differentiation between abstracts and RAs as text types.

Additionally, the data also indicate that there is variation in tenor of discourse, specially in the sub-parameter of *social role relationship* in which the feature *modals* play a very important role as an indicator for this sub-parameter (cf. Table 4.3). Finally, there is a minor variation also in the mode of discourse, especially due to the variation in its indicator lexical density (cf. Table 4.3).

The next chapter, Chapter 6, will discuss the overall results under a broader perspective, aiming to get the whole picture of relationship between abstracts and their research articles. Chapter 6 therefore addresses the main research questions and hypotheses formulated in Section 4.3 in order to corroborate or refute them based on the overall results of this research.

Chapter 6

Conclusions

This last chapter of this thesis presents a summary of the findings (Section 6.1), followed by a discussion of the results (Section 6.2) and of the methodology used (Section 6.3). Finally, Section 6.4 addresses some research aspects for future work.

6.1 Summary

The first chapter of this thesis, Chapter 1, discussed the motivations leading to the research questions and goals of this study. Primarily, this research aims to investigate linguistic differences between abstracts and their research articles based on the quantitative distribution of selected linguistic features at both lexical and grammatical levels, and to explore the relationship between these two text types in a broader linguistic context.

The following chapter, Chapter 2, was dedicated to the state-of-the-art and presented an overview of the linguistic research on abstracts and research articles including their historical development. It discussed the most relevant approaches adopted in current linguistic research such as genre analysis (e.g., Swales 1990, 2004), the area of linguistics where most of the studies concerning abstracts and research articles are to be placed; register analysis, an approach for investigation of linguistic variation, mainly represented by Douglas Biber's work (e.g., Biber 1988, 1995); and the theoretical underpinnings of this research, Systemic Functional Linguistics (Halliday 1985a, 2004a), a sophisticated linguistic model. Systemic Functional Linguistics makes possible the analysis of the relations between language and different social contexts and allows a detailed investigation of discourse variation based on the analysis of concrete linguistic features. Finally, the controversies involving the concepts of genre and register were addressed in connection with the linguistic model proposed by the theoretical underpinnings of this research.

Chapter 3 discussed methodological aspects in current linguistic research, setting the scene for the methodology adopted in this research. The approach adopted in this research was more quantitative than qualitative. In order to characterize abstracts and research articles linguistically, linguistic features were identified, quantified and analyzed. For this reason, Chapter 3 comprised a brief overview on empiricism in linguistics and an analysis of the empirical methodology adopted precisely including an exploration of its advantages and disadvantages. Moreover, the synergies and adequacies between corpus linguistics, as a methodology, and Systemic Functional Linguistics, as theoretical background for linguistic research on language variation were exploited. The issue of statistical evaluation of linguistic data was also addressed, since statistics is fundamental in the process of description, visualization, evaluation and interpretation of quantitative results in linguistics, including not only its advantages but also its limitations.

Chapter 4 described the concrete design of this research. The corpus under study, AbstRA, was introduced, followed by the description of its processing and annotation steps. The hypotheses to be tested in the empirical analysis were then formulated and the question of the choice of the linguistic features for the empirical analysis was addressed. Finally, the features chosen for the quantitative analysis were presented together with a discussion of their relationship as adequate indicators of the parameters for the linguistic description of language according to the linguistic model used as theoretical background of this research, Systemic Functional Linguistics.

Chapter 5 presented all the results obtained in this research, which followed a twofold empirical analysis plan. First, a deductive empirical analysis was performed, by which selected features were quantitatively determined and statistically evaluated for significance and hypotheses testing. The linguistic features chosen for the deductive empirical analysis were: sentence length, type/token ratio, lexical words (nouns, adjectives, adverbs, verbs), lexical density, most frequent lexical items, keywords, modals, passive voice, and grammatical complexity. Abstracts showed statistically significant higher frequencies of all these features in comparison to their research articles apart from modals, which results indicated that abstracts use modals significantly less frequently than RAs, and from lexical density, which results were slightly below the border of significancy. Thus, abstracts and research articles are statistically different from each other.

According to Table 4.3 (cf. p. 67), the linguistic features chosen for the quantitative analysis are indicative of the parameters of context of situation: field, tenor and mode of discourse. Therefore, a statistically significant difference for a given feature between abstracts and research articles implies differences in the configuration of the parameter of context of situation for which this given feature is indicative of. For instance, the fact that there is a statistically significant difference between abstracts and research articles for the feature *nominalization* can be interpreted as indicative of differences in the configuration of the parameters of context of situation field of discourse (subcategory: goal orientation) and tenor of discourse (subcategory: social role relationship), as described in Table 4.3. Following this rationale for all the chosen features, it can be inferred that the results indicated significant variation between abstracts and research articles in all three parameters of context of situation, i.e., field, tenor and mode of discourse, since abstracts showed statistically significant results for all these features in comparison to their research articles (apart from lexical density).

The results also revealed that abstracts and research articles are very distinctive types of texts. Abstracts showed significantly higher frequencies of occurrences of features which are very typical indicators of expository texts, such as nouns, lexical density, sentence length, and nominalizations. In contrast, the lower frequency of these same and also other features in research articles, such as modals and adverbs, can be interpreted as indicative of properties typical of argumentative texts (cf. Section 4.4).

Finally, the data showed a statistically significant difference across domains, i.e., computer science, linguistics, biology and mechanical engineering, within abstracts and within research articles for the chosen linguistic features.

In relation to the main hypotheses tested in this research, which were formulated in Chapter 4, the data obtained from analysis refuted all the three null hypotheses ($\mathbf{H}_0$). Hence, all three alternative hypotheses were adopted in this research (cf. Section 3.4.2, p. 44), to wit:

- The quantitative analysis of linguistic features revealed statistically significant differences between abstracts and their research articles at both lexical and grammatical levels (**H1**).
- The quantitative analysis of linguistic features revealed statistically significant differences across domains for abstracts and research articles at both lexical and grammatical levels (**H2**).

- Abstracts and their research articles showed different configurations of the parameters of context of situation field, tenor, and mode of discourse (**H3**).

The second part of the empirical analysis plan of this research comprised an inductive empirical analysis. The purpose of the inductive empirical analysis was to corroborate the results of the deductive empirical analysis without hypotheses formulation prior to the experiments. The features chosen for the inductive empirical analysis were: text type (abstracts or RAs), domain, prepositions, adjectives, modals, nouns, personal pronouns, possessive pronouns, adverbs, present tense, past participle, past tense, nominalizations, sentence length and lexical density (cf. Section 5.2). Since the results obtained in such a "theory-free" analysis like the inductive methods adopted here corroborated the data obtained in the descriptive analysis, they also confirmed the adequacy of the hypotheses and features chosen for the deductive empirical analysis.

The methods for inductive empirical analysis adopted in this research, hierarchical agglomerative cluster analysis and principal component analysis, allowed the surfacing of additional insight into similarities and dissimilarities of the texts of the AbstRA corpus. The hierarchical agglomerative cluster analysis showed that the texts were grouped into two main clusters, abstracts and research articles, with just a few mismatches (cf. Figure 5.18, p. 145). It was also observed that the texts of different disciplines, i.e., linguistics, computer science, biology and mechanical engineering, tend to be grouped per discipline both within the cluster of abstracts as well as in the cluster of the research articles. This is an indication of domain specific variation in the AbstRA corpus.

The principal component analysis indicated that the first principal component with high positive loadings of modals, personal pronouns, possessive pronouns, adverbs, and present tense matched precisely one of the main clusters obtained by the hierarchical agglomerative cluster analysis (cf. Figure 5.23, p. 151 and Section 5.2.2, p. 153). The features in this cluster are very typical for characterizing argumentative discourse. This first principal component also showed high negative loadings of nouns, past tense, lexical density, sentence length, and nominalizations which matched with the other main cluster obtained by the hierarchical agglomerative cluster analysis (cf. Figure 5.23, p. 151). These features are typical for expository discourse. The second principal component had lexical density as the main positively loaded feature, and several highly negative loaded features such as prepositions, nominalizations, past participle, and adjectives. The positive loadings of nouns, prepositions, adjectives, etc., indicated a very informational discourse and careful integration of information in text. In contrast, the high negative loadings was associated with a non-abstract style. The third principal component had high positive loadings for the features sentence length, past tense, personal pronouns and possessive pronouns and high negative loadings for lexical density, modals, nouns, and present tense which represent not overtly argumentative texts (Conrad & Biber 2001: 36).

The bi- and tridimensional visualization of the results of the inductive analysis supported the interpretation that the texts of the AbstRA corpus are clearly grouped according to their text type, i.e., whether they are abstracts or research articles (cf. Figures 5.27, 5.28 and 5.29, p. 159 - 161). According to the data obtained, it can be concluded that abstracts and RAs are distinct text types from each other. Furthermore, it was graphically shown that there is a distinction of the texts of the corpus according to their domains (cf. Figures 5.30, 5.31 and 5.32, p. 162 - 164).

Overall, the data obtained in the inductive empirical analysis corroborated the former inference derived from the results of the deductive empirical analysis that abstracts are more expository texts, with a high informational and abstract discourse. In contrast, research articles showed more properties of argumentative and persuasive texts indicating a greater involvement of the author in the text than in the abstracts.

Therefore, the choice of the linguistic features for the empirical quantitative analysis proved to be adequate in addressing the initially proposed research questions using the methodology discussed in Chapters 3 and 4. The aims of this research (cf. p. 4 and 167) were thus achieved by developing a framework to identify and quantitatively evaluate linguistic differences between abstracts and their research articles both at the grammatical and lexical levels.

6.2 Discussion of results

This section addresses the issue of how abstracts and research articles are related in a broader linguistic context. It aims to answer the question whether abstracts and research articles are different registers or different genres (and in which regard), based on the results obtained in this research.

According to the theoretical background of this research, Systemic Functional Linguistics, language and context are “inextricably linked” (Thompson 2004: 10). Systemic Functional Linguistics defines *register* as a pattern of language according to *use* in context. As discussed in Chapter 2, language model of Systemic Functional Linguistics is composed of three metafunctions: ideational, interpersonal, and textual. The configurations of these are determined by the situational context in which language is used. These three metafunctions are expressed through the three parameters of the context of situation, field, tenor and mode of discourse, respectively. Language variation thus reflects on a variation in the configurations of field, tenor, and mode (cf. Figure 2.3, p. 27). Therefore, these parameters are considered adequate parameters for the study of language variation, i.e., *register* variation (cf. e.g., Halliday & Hasan 1989). Chapter 4 addressed the issue that different configurations of field, tenor, and mode of discourse, are directly reflected in different realizations of concrete linguistic features – or indicators – which can be used for quantitative empirical analysis. However, the relationship between each feature as an indicator of a parameter of context of situation is *not* one-to-one, but many-to-many since one feature can be an indicator of more than one parameter (cf. Section 4.4). As mentioned in Section 6.1, statistically significant results for a certain feature imply significant differences in the configuration of all parameters of context of situation for which this given feature is indicative of. The results for all linguistic features showed statistically significant differences (apart from lexical density). These differences were found not only in the comparison between abstracts and research articles but also in the comparison across the four different domains within abstracts and within research articles: linguistics, computer science, biology and mechanical engineering. Furthermore, both inductive and deductive approaches lead to similar results and interpretations. Since these linguistic features chosen for the analysis here are acknowledged to be indicators of the parameters of field, tenor, and mode of discourse, it can be inferred that there are significant differences in the configurations of theses parameters. Hence, it can be concluded that there is *register* variation, i.e., language variation according to use, between abstracts and research articles.

The first results to be considered are those from the descriptive analysis. Lexical words, modals, nominalizations, passives, and type/token ratio are linguistic features, which are indicative of the sub-category goal orientation of the field of discourse (cf. Table 4.3, p. 67). Since all these indicators showed statistically significant differences between abstracts and research articles, it can be concluded that there is a significant difference in the configuration of the goal orientation between abstracts and research articles. As mentioned in Section 6.1, the results of the deductive empirical analysis showed that abstracts are more *expository* texts, with a high informational and abstract discourse. In contrast, research articles showed properties of *argumentative* and persuasive texts (cf. p. 153). The domain specific variation observed in the data is interpreted as a consequence of variation in the field of discourse, more specifically in the experiential domain. This subparameter of the field of discourse, refers to what is happening, to the topic and to the nature of the action taking place in the discourse. The indicators of experiential domain are the features keywords and lexical words, specially nouns. These features also showed significant differences between abstracts and research articles, specially across disciplines. For this reason, it can be concluded that there is significant difference in the experiential domain.

The results also indicated significant differences between abstracts and research articles in the configurations of the tenor of discourse and its subcategories, social role relationship and social distance, indicators of which are modals, nominalizations, sentence length, lexical words, type/token ratio, and grammatical complexity (cf. Table 4.3, p. 67). A typical feature characterizing the level of authority, a subcategory of the social role relationship, is the use of modality. The results showed that abstracts use significantly less modals than research articles. This can be interpreted as a consequence of authors trying to achieve a more distant, powerful and persuasive social role in the communication process with the readers in the research articles. The results concerning the use of nominalizations and sentence length are an additional indication that there are significant differences at the level of expertise, another subcategory of social role relationship, between abstracts and research articles. Finally, as mentioned in Section 4.4, social distance is concerned with the level of formality in the communication taking place. This parameter is similar to Biber's dimension "involved vs informational production". The results obtained showed that abstracts use lexical words significantly more frequently and have significantly higher type/token ratio than research articles. This can be interpreted as an indication that abstracts tend to be more informational than research articles.

For the last parameter of context of situation, mode of discourse and its two sub-categories, language role and medium, there is not such a clear distinction between abstracts and RAs. The main reason for this is due to the results for the indicator lexical density, which are not significant, yet almost at the border of significance. However, there is a minor variation in the subcategory of the mode of discourse – medium – due to the variation in its indicator grammatical complexity.

Hence, it can be affirmed that there is *register* variation between abstracts and research articles, specially across disciplines, as a consequence of variation in language use, i.e., variation in the *function* of language in the context of situation where communication takes place.

It should be borne in mind, that the expected variation between abstracts and their research articles *per se* is narrower than the variation expected between other text types, like for instance, news in comparison to scientific discourse. This is because abstracts and their research articles are intrinsically much more similar to each other than to other text types. According to the continuous axis for discourse variation proposed by Biber (e.g., 1988, 1995), abstracts and their research articles are to be placed very close to each other. Still, the differences found in this research showed a more distinctive variation compared to the initial assumptions.

The last issue to be considered here is whether abstracts and research articles show *genre* variation. As discussed in Section 2.5, the genre definitions adopted in this research followed those proposed by Martin (1992a: 505-507). Martin considers register as patterns of linguistic patterns and *genre* as patterns of register patterns. Moreover, in consonance with Martin, *genre* is also defined as a “staged, *goal-oriented* social process realized through register” (Martin 1992a: 505; emphasis added). Accordingly, it can be said that genre is concerned with the *purpose* of discourse, while register is concerned with the *function* of discourse.

As discussed in Section 6.1, the results indicated a domain specific variation of discourse between articles and their research articles. This phenomenon complies with the initial assumptions in this research and with former studies (cf. Section 2). It can be interpreted as being a reflection of the different *functions* of discourse in different disciplinary contexts, i.e., different *registers*.

However, more importantly, the data obtained in this research showed a clear distinction between the linguistic properties of abstracts and their research articles, specially those related to the parameter *goal orientation*, which is related to the *purpose* of the discourse in a given context of situation. As mentioned previously, abstracts showed to be *expository* texts, with a high informational and abstract discourse, while their research articles showed properties of *argumentative* texts. Therefore, based on Martin’s definitions of *genre*, the statistically significant differences in the *goal orientation* of abstracts and research articles can be interpreted as being related to different *purposes* of their discourses in the context of situation. For this reason, it can be affirmed that the data obtained in this research contradicted the postulates of Swales (1990); Hyland (2004) and Swales & Feak (2009), who consider abstracts as a “part-genre” of research articles. Furthermore, although being physically contained in research articles, the abstracts under study are all *stand-alone* complete texts; and this is also an important criteria for defining *genre* (cf. Biber & Conrad 2009: 33). Finally, taking into consideration former studies that exemplarily showed that the *purpose* of abstracts goes beyond the mere summarization of the research article (cf. Section 2, e.g., Hyland 2004: 63-65), it can be said that the results of this research substantiate the contention that abstracts and their research articles in the corpus under study are distinct *genres*, although the former is physically embedded in the latter.

6.3 Brief assessment of the methodology

Overall, the methodology adopted in this research lead to relevant and interesting results and allowed the investigation of the proposed research questions. It proved to be adequate for testing the formulated hypotheses thereby revealing important linguist characteristics of the objects of study, abstracts and their research articles. The methodology used here can be summarized as follows:

- it is in accordance with the theoretical background adopted in this research since it established a clear link between the theoretical framework and the concrete linguistic features chosen for the empirical analysis;
- it complies with current practices in corpus linguistics;
- it addresses both deductive and inductive approaches of empirical quantitative analysis which supported and complemented each other in the process of data interpretation;
- it uses several statistical techniques in order to determine the significance of the results and to support their interpretation.

However, the methodology has also its limitations. For instance, due to the relative small size of the corpus under study, other inductive techniques could not be applied, e.g., factor analysis, which would have potentially delivered better and even more reliable results. Moreover, the investigation of additional disciplines and linguistic features would have enriched the profiling of the discourse in abstracts and research articles. Since a corpus represents a given language only partially, studies on larger corpora can probably deliver better results in comparison to smaller corpora. However, independent of the its size, a corpus will never represent the wholeness of language. For this reason, the data obtained in a corpus analysis primarily represent the properties of the studied corpus.

Another important issue on the methodology used in this research is that all statistical techniques are based on probabilities and approximations. Consequently, statistical results generally only represent likeness and tendencies, not the absolute trueness of facts. As is the case for all statistical analyses, the results obtained here allow only statements about probabilities. Still, the methodology applied here is the best suitable methodology to date. Since the size and composition of the AbstRA corpus comply with the requirements in current corpus linguistic methodology (cf. Section 4.1), the validity and repeatability of the results are assured. The methodological issues addressed here should be seen as a motivation for further studies involving the investigation of linguistic properties of scientific discourse, as addressed in the next and closing section of this thesis.

6.4 Future work

In this study, a set of lexico-grammatical features was quantitatively analyzed over a synchronic corpus of abstracts and research articles aiming to gain insight into the linguistic characteristics of these texts supported by solid theoretical underpinnings and statistical evaluation of the results. However, this studied did not cover all possible perspectives concerning the investigation of language variation.

One main issue that was not addressed in this research is the linguistic evolution of such texts over time. A main further path to be explored would be the investigation of such lexico-grammatical features over a diachronic corpus of abstracts and research articles, in order to investigate how the linguistic characteristics of these texts developed and changed within a span of time. Furthermore, a comparison between abstracts, research articles and other scientific discourses, which also imply building a bigger corpus, would allow a better understanding of social and cultural practices in academia that are reflected in their discourse.

Although the number of linguistic features studied and the data obtained provided enough evidence for corroborating the hypothesis that abstracts and research articles are linguistically quite different from each other, the inclusion of more features in a future research would lead to a wider profile of the linguistic characteristics of these genres. By enlarging the number of features used in a similar quantitative analysis, other statistical techniques for evaluation of the results become usable, which in turn would allow a better data interpretation. Furthermore, the examination of the relationship between such additional features as indicators of the configuration parameters within a complex language model, like the one adopted here, would contribute for more detailed theoretical mapping of language variation.

Finally, this research can be seen as a further small gravel in the long pathway of linguistic investigation of scientific discourse. Not only linguists interested in studies on language variation can profit from the approach and methodology employed and the results obtained here. A study of the linguistic properties of scientific discourse like this one can contribute to the area of English for Special Purposes, having pedagogical applications in teaching of contemporary academic and research English inasmuch as understanding a certain discipline and practices of its community involves understanding their literacy.

References

Atkinson, D. (1992). The Evolution of Medical Research Writing from 1735 to 1985: The Case of the Edinburgh Medical Journal. *Applied Linguistics*, *13*(4), 337–374.

Baayen, R. H. (2008). *Analysing Linguistic Data. A Practical Introduction to Statistics using R.* London [u.a.]: Cambridge University Press.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In G. F. Baker M. & E. Tognini-Bonelli (Eds.) *Text and Technology: in Honour of John Sinclair*, (pp. 233–250). Amsterdam: Benjamins.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target*, *7*(2), 223–243.

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. L. Somers (Ed.) *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, (pp. 175–186). Amsterdam: Benjamins.

Banks, D. (1991). Some Observations Concerning Transitivity and Modality in Scientific Writing. *Language Sciences*, *13*(1), 59–78.

Banks, D. (1994). Clause organization in the scientific journal article. *ALSED-LSP*, *17*(38), 4–16.

Banks, D. (2005a). Emerging scientific discourse in the late seventeenth century. A comparison of Newton's Optiks, and Huygens' Traite de la lumiere. *Functions of Language*, *12*(1), 65–86.

Banks, D. (2005b). On the historical origins of nominalized process in scientific text. *English for Specific Purposes*, *24*(3), 347–357.

Banks, D. (2006). Referring to others in the scientific journal article. A brief history. *Linguistics and the Human Sciences*, *2*(3), 329–353.

Banks, D. (2008). *The Development of Scientific Writing. Linguistic Features and Historical Context*. London, Oakville: Equinox.

Baroni, M. & Evert, S. (2008). Statistical methods for corpus exploitation. In A. Lüdeling & M. Kytö (Eds.) *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.

Bawarshi, A. S. & Reiff, M. J. (2010). *Genre: An Introduction to History, Theory, Research, and Pedagogy*. West Lafayette, Indiana: Parlor Press.

Bazerman, C. (1984a). Modern evolution of the experimental report in physics: spectrocospic articles in *Physical Review*, 1893-1980. *Social Studies in Science*, *14*(2), 163–96.

Bazerman, C. (1984b). The Writing of Scientific Non-Fiction: Contexts, Choices and Constraints. *Pre/Text*, *5*(1), 39–74. Reprinted in Ten Years of Pre/Text, Vitanza, V. (Ed.), University of Pittsburgh Press, 1994.

Bazerman, C. (1988). *Shaping written knowledge: the genre and activity of the experimental article in science*. Madison, Wisconsin: University of Wisconsin Press.

Bazerman, C. (1994). Systems of genres and the enactments of social intentions. In A. Freedman & P. Medway (Eds.) *Genre and the new rhetoric*, (pp. 79–101). London: Taylor & Frances.

Beaugrande, R. d. (1993). 'Register' in discourse studies: a concept in search of a theory. In M. Ghadessy (Ed.) *Register Analysis. Theory and Practice*, (pp. 7–25). London and New York: Pinter.

Bhatia, V. K. (1993). *Analysing genre: language use in professional settings*. London: Longman.

Bhatia, V. K. (2002). A generic view of academic discourse. In J. Flowerdew (Ed.) *Academic Discourse*, (pp. 21–39). Harlow: Longman.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, *5*, 257–269.

Biber, D. (1993a). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*(4), 243–257.

Biber, D. (1993b). The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings. *Computers and the Humanities*, *26*, 331–345.

Biber, D. (1994). An Analytical Framework for Register Studies. In D. Biber & E. Finegan (Eds.) *Sociolinguistic Perspectives on Register*, (pp. 31–56). New York, Oxford: Oxford University Press.

Biber, D. (1995). *Dimensions of Register Variation. A cross linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D. (1996). Investigating Language Use Through Corpus-Based Analysis of Association Patterns. *International journal of corpus linguistics*, *1*(2), 171–197.

Biber, D. (2006a). Analytical procedures for the linguistic analyses. In *University Language: A Corpus-based Study of Spoken And Written Registers*, (pp. 241–250). Amsterdam/Philadelphia: John Benjamins Publishing.

Biber, D. (2006b). Methodological issues in quantitative vocabulary analyses. In *University Language: A Corpus-based Study of Spoken And Written Registers*, (pp. 251–257). Amsterdam/Philadelphia: John Benjamins Publishing.

Biber, D. (2006c). Multi-dimensional patterns of variation among university registers. In *University Language: A Corpus-based Study of Spoken And Written Registers*, (pp. 177–212). Amsterdam/Philadelphia: John Benjamins Publishing.

Biber, D. (2006d). *University Language: A Corpus-based Study of Spoken And Written Registers*. Amsterdam/Philadelphia: John Benjamins Publishing.

Biber, D., Connor, U. & Upton, T. A. (2007). *Discourse on the move: using corpus analysis to describe discourse structure*. Amsterdam: John Benjamins Publishing.

Biber, D. & Conrad, S. (2009). *Register, Genre and Style*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S. & Leech, G. (2002). *Longman Student Grammar of Spoken and Written English*. Harlow: Longman.

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, D. & Finegan, E. (1989). Drift and the evolution of English style: a history of three genres. *Language*, *65*(3), 487–517.

Biber, D. & Finegan, E. (1992). The linguistic evolution of five written and speech-based English genres from the 17th to the 20th centuries. In M. Rissanen, O. Ihalainen, T. Nevalainen & I. Taavitsainen (Eds.) *History of Englishes: New Methods and Interpretations in Historical Linguistics*, (pp. 688–704). Mouton de Gruyter.

Biber, D. & Finegan, E. (Eds.) (1994). *Sociolinguistic Perspectives on Register*. New York, Oxford: Oxford University Press.

Biber, D., Johansson, S. & Leech, G. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Bondi, M. (2004). The discourse function of contrastive connectors in academic abstracts. In K. Aijmer & A.-B. Stenström (Eds.) *Discourse Patterns in Spoken and Written Corpora*, Pragmatics & Beyond: New Series (P&B): 120, (pp. 139–156). Amsterdam, Netherlands: Benjamins.

Bondi, M. & Scott, M. (Eds.) (2010). *Keyness in Texts*. London: John Benjamins Publishing.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation: für Human- und Sozialwissenschaftler*. Berlin: Springer, 4. ed.

Busch-Lauer, I.-A. (1995). Textual organization in English and German medical abstracts. In B. Warvik, S.-K. Tanskanen & R. Hiltunen (Eds.) *Organization in Discourse.*, vol. 14 of *Anglicana Turkuensia*. Turku: Univ. of Turku. Proc. from Turku Conf.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1962). A transformational approach to syntax. In A. A. Hill (Ed.) *Proceedings of the 3rd Texas Conference on Problems of Linguistic Analysis in English (1958)*. Austin: University of Texas Press.

Chomsky, N. (1964). Formal discussion: the development of grammar in child language. In U. Bellugi & R. Brown (Eds.) *The Acquisition of Language*, (pp. 37–39). Indiana: Purdue University.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: M.I.T. Press.

Chomsky, N. (1975). *The Logical Structure of Linguistic Theory*. New York: Plenum Press.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. (1984). *Modular Approaches to the Study of the Mind*. San Diego University Press.

Chomsky, N. (1988). *Generative Grammar: Its Basis, Development and Prospects*. Kyoto: Kyoto University of Foreign Studies.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, London: The MIT Press.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proceedings of the 3rd Conference on Computational Lexicography and Text research (COMPLEX 94)*, (pp. 23–32). Budapest, Hungary.

Christ, O., Schulze, B. M., Hofmann, A. & König, E. (1999). The IMS Corpus Workbench: Corpus Query Processor (CQP) User's Manual. Tech. rep., University of Stuttgart, Institute for Natural Language Processing.

Conrad, S. (1996). Investigating academic texts with corpus-based techniques: an example from biology. *Linguistics and Education*, *8*, 299–326.

Conrad, S. & Biber, D. (Eds.) (2001). *Variation in English: Multi-Dimensional Studies*. London: Longman.

Crawley, M. J. (2007). *The R Book*. UK: Wiley and Sons, Ltd.

Dorgeloh, H. & Wanner, A. (2003). Too abstract for agents? The syntax and semantics of agentivity in abstracts of English research articles. In W. Bisang, H. H. Hock & W. Winter (Eds.) *Mediating between Concepts and Grammar*, Trends in Linguistics. Studies and Monographs 152, (pp. 433–455). Mouton de Gruyter.

Eckart, R. (2006). *A Framework For Storing, Managing and Querying Multi-Layer Annotated Corpora*. Diplomarbeit, Technische Universität Darmstadt, Darmstadt.

Eckart, R. & Teich, E. (2007). An XML-based data model for flexible representation and query of linguistically interpreted corpora. In G. Rehm, A. Witt & L. Lemnitzer (Eds.) *Data Structures for Linguistic Resources and Applications – Proceedings of the Biennial GLDV Conference 2007*, (pp. 327–336). Tübingen, Germany: Gunter Narr Verlag.

Fillmore, C. (1992). Corpus linguistics or computer-aided armchair linguistics. In J. Svartvik (Ed.) *Directions in Corpus Linguistic, Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. ACM, Berlin, New York: Mouton de Gruyter.

Firth, J. R. (1957). *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Firth, J. R. (1968). *Selected Papers of J. R. Firth 1952-1959*. London: Longmans.

Fluck, H.-R. (1988). Zur Analyse und Vermittlung der Textsorte 'Abstract'. In C. Gnutzmann (Ed.) *Fachbezogener Fremdspracheunterricht*, (pp. 67–90). Tübingen, Germany: Gunter Narr Verlag Tübingen.

Gerbert, M. (1970). *Besonderheiten der Syntax in der Technischen Fachsprache des Englishen*. Berlin: Halle.

Ghadessy, M. (Ed.) (1988). *Registers of Written English: Situational Factors and Linguistic Features*. London: Pinter.

Ghadessy, M. (Ed.) (1993). *Register Analysis. Theory and Practice*. London: Pinter.

Ghadessy, M. (1999). Textual Features and Contextual Factors for Register Indentification. In M. Ghadessy (Ed.) *Text and Context in Functional Linguistics*, (pp. 125–139). Amsterdam, Philadelphia: John Benjamins B.V.

Ghadessy, M. (2003). Comments on Douglas Biber, Susan Conrad, Randi Reppen, Pat Byrd, and Marie Helt's 'Speaking and Writing in the University: A Multidimensional Comparison'. A Reader reacts... *TESOL Quarterly*, *37*(1), 147–150.

Gilquin, G. & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, *5*(1), 1–26.

Gledhill, C. (2000a). The discourse function of collocation in research article introductions. *English for Specific Purposes*, *19*(2), 115–135.

Gledhill, C. J. (2000b). *Collocations in Science Writing*. Tübingen: Gunter Narr Verlag.

Gnutzmann, C. (1991). 'Abstracts' und 'Zusammenfassungen' im deutsch-englischen Vergleich: Das Passiv als interkulturelles und teiltextdifferenzierendes Signal. In B.-D. Müller (Ed.) *Interkulturelle Wirtschaftskommunikation*, (pp. 363–378). Iudicum Verlag.

Graetz, N. (1982). Teaching EFL Students to Extract Structural Information from Abstracts. *Paper presented at the International Symposium on Language for Special Purposes (Eindhoven, The Netherlands, August 2-4, 1982)*, (pp. 1–23).
URL `http://www.eric.ed.gov:80/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/2f/f1/a9.pdf`

Gregory, M. (1967). Aspects of varieties differentiation. *Journal of Linguistics*, *3*(2), 177–198.

Gries, S. T. (2006). Exploring variability within and between corpora: some methodological considerations. *Corpora*, *1*(2), 109–151.

Gries, S. T. (2007). Cognitive Linguistics and Functional Linguistics: Common assumptions and methods. In S. T. Gries & A. Stefanowitsch (Eds.) *Corpora in Cognitive Linguistics. Corpus-Based Approaches to Syntax and Lexis*, (pp. 1–17). Berlin, New York: Mouton de Gruyter.

Gries, S. T. (2008a). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*(4), 403–437.

Gries, S. T. (2008b). *Statistik für Sprachwissenschaftler*. Göttingen: Vandenhoeck & Ruprecht.

Gries, S. T. (2009a). *Quantitative Corpus Linguistics with R. A practical introduction*. New York and London: Routledge.

Gries, S. T. (2009b). *Statistics for Linguistics with R. A Practical Introduction*. Berlin, New York: Mouton de Gruyter.

Gries, S. T. (2010). Corpus linguistics and theoretical linguistics. A love-hate relationship? Not necessarily... *International Journal of Corpus Linguistics*, *15*(3), 327–343.

Gustafsson, L. O. (2006). The passive in nineteenth-century scientific writing. In M. Kytö, M. Rydén & E. Smitterberg (Eds.) *Nineteenth-century English*, (pp. 110–135). New York: Cambridge University Press.

Haegeman, J. (1991). *Introduction to Government and Binding Theory*. Oxford: Blackwell.

Halliday, M. A. K. (1959). The secret history of the mongols. In *Studies in Chinese Language, Eighth volume in the collected works of M.A.K. Halliday (Reprinted)*, (pp. 5–171). London: Continuum.

Halliday, M. A. K. (1978). *Language as Social Semiotic*. London: Edward Arnold.

Halliday, M. A. K. (1985a). *An Introduction to Functional Grammar*. London: Arnold.

Halliday, M. A. K. (1985b). *Spoken and Written Language*. Victoria: Deakin University.

Halliday, M. A. K. (1988). On the language of physical science. In M. Ghadessy (Ed.) *Registers of Written English: Situational Factors and Linguistic Features*, (pp. 162–177). London: Pinter.

Halliday, M. A. K. (1992). Language as system and language as instance: The Corpus as a theoretical construct. In J. Svartvik (Ed.) *Directions in Corpus Linguistic, Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. ACM, Berlin, New York: Mouton de Gruyter.

Halliday, M. A. K. (1993a). On the language of Physical Science. In M. A. K. Halliday & J. R. Martin (Eds.) *Writing Science: Literacy and Discursive Power*, (pp. 54–68). London, Washington D.C.: The Falmer Press.

Halliday, M. A. K. (1993b). Some Grammatical Problems in Scientific English. In M. A. K. Halliday & J. Martin (Eds.) *Writing Science: Literacy and Discursive Power*. London: University of Pittsburgh Press.

Halliday, M. A. K. (2004a). *An Introduction to Functional Grammar*. Revised by C. M. I. M. Matthiessen. London: Arnold, 3 ed.

Halliday, M. A. K. (2004b). The Language of Science. In J. J. Webster (Ed.) *Collected Works of M. A. K. Halliday*, vol. 5. London, New York: Continuum.

Halliday, M. A. K. (2004c). Things and Relations: Regrammaticizing Experience as Technical Knowledge. In J. J. Webster (Ed.) *The Language of Science*, (pp. 49–101). London: Continuum.

Halliday, M. A. K. (2006). Afterwords. In G. Thompson & S. Hunston (Eds.) *System and Corpus: Exploring connections*, (pp. 293–299). London: Equinox.

Halliday, M. A. K. (2008). *Complementarities in Language*. Beijing: The Commercial Press.

Halliday, M. A. K. (2009). Methods – techniques – problems. In M. A. K. Halliday & J. Webster (Eds.) *Continuum Companion to Systemic Functional Linguistic*, (pp. 59–86). London: Continuum.

Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Halliday, M. A. K. & Hasan, R. (1989). *Language, context and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Halliday, M. A. K. & Martin, J. R. (1993). *Writing Science: Literacy and Discursive Power*. London, Washington D.C.: The Falmer Press.

Halliday, M. A. K. & Matthiessen, C. M. I. M. (2006). *Construing Experience Through Meaning. A Language-based Approach to Cognition*. London, New York: Continuum.

Halliday, M. A. K., McIntosh, A. & Strevens, P. (1964). *The Linguistic Sciences and Language Teaching*. London: Longman.

Hartley, J. (1999). Applying ergonomics to Applied Ergonomics: using structured abstracts. *Applied Ergonomics*, *30*(6), 535–541.

Hartley, J. & Sydes, M. (1997). Are structured abstracts easier to read than traditional ones? *Journal of Research in Reading*, *20*(2), 122–136.

Hartley, J., Sydes, M. & Blurton, A. (1996). Obtaining information accurately and quickly: are structured abstracts more efficient? *Journal of Information Science*, *22*(5), 349–356.

Hoey, M. (2005). *Lexical Priming*. London, New York: Routledge.

Horrocks, G. (1987). *Generative Grammar*. London: Longman.

Hunston, S. (1993). Evaluation and ideology in scientific writing. In M. Ghadessy (Ed.) *Register Analysis. Theory and Practice*, (pp. 57–73). London, New York: Pinter Publishers.

Hyland, K. (1998). *Hedging in scientific research articles*. Amsterdam: John Benjamins.

Hyland, K. (1999). Talking to students: metadiscourse in introductory coursebooks. *English for Specific Purposes*, *18*(1), 3–26.

Hyland, K. (2002). *Teaching and researching writing*. Harlow: Pearson Education.

Hyland, K. (2003). Genre-based pedagogies: A social response to process. *Journal of Second Language Writing*, *12*, 17–29.

Hyland, K. (2004). *Disciplinary Discourses. Social Interactions in Academic Writing*. Michigan Classics Edition. Ann Arbor: The University of Michigan Press. First published by Pearson Education Limited, Longman, 2000.

Hyland, K. (2008). Academic clusters: text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, *18*(1), 41–62.

Hyland, K. (2009). *Academic Discourse*. Continuum Discourse Series. London, New York: Continuum.

Hymes, D. (1974). *Foundations in sociolinguistics: an ethnographic approach*. Philadelphia: University of Pennsylvania Press.

Inman, M. (1978). Lexical analysis of scientific and technical prose. In T. Trimble, M. Trimble & K. Drobnic (Eds.) *English for specific purposes: Science and technology*, (pp. 242–256). Corvallis OR: English Language Institute, Oregon State University.

Jackendoff, R. (1974). *Semantic Interpretation in Generative Grammar*. Cambridge, London: The MIT Press.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer New York.

Jordan, M. P. (1991). The linguistic genre of abstracts. In A. Della Volpe (Ed.) *The 17th LACUS forum 1990*, (pp. 507–527). Lake Bluff: IL: LACUS.

Kaplan, R., Cantor, S., Hagstrom, C., Lia, D., Shiotani, Y. & Zimmerman, C. (1994). On abstract writing. *Text*, *14*(3), 401–426.

Klein, D. & Manning, C. D. (2003a). Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, (pp. 423–430).

Klein, D. & Manning, C. D. (2003b). Fast Exact Inference with a Factored Model for Natural Language Parsing. *Advances in Neural Information Processing Systems (NIPS 2002)*, *15*, 3–10.

Kline, P. (1994). *An easy guide to factor analysis*. London: Routledge.

Kretzenbacher, H. L. (1990). *Rekapitulation. Textstrategien der Zusammenfassung von wissenschaftlichen Fachtexten*. Tübingen, Germany: Gunter Narr Verlag Tübingen.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.) *Directions in Corpus Linguistic, Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. ACM, Berlin, New York: Mouton de Gruyter.

Liddy, E. D. (1991). The discourse-level structure of empirical abstracts: an exploratory study. *Information Processing & Management*, *27*(1), 55–81.

Lorés, R. (2004). On RA abstracts: from rhetorical structure to thematic organisation. *English for Specific Purposes*, *23*(3), 280–302.

Love, A. (1993). Lexico-semantic features of geology textbooks. *English for Specific Purposes*, *12*(3), 197–218.

Luke, A. (1996). Genres of Power? Literacy Education and the Production of Capital. In R. Hasan & G. Williams (Eds.) *Literacy in Society*, (pp. 308–338). London: Longman.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Marco, M. J. L. (2000). Collocational frameworks in medical research papers: a genre-based study. *English for Specific Purposes*, *19*(1), 63–86.

Marcus, M. P., Santorini, B. & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Martin, J. R. (1983). The development of register. In J. Fine & R. O. Freedle (Eds.) *Development Issues in Discourse*. Norwood, New Jersey: Ablex Publishing Corporation.

Martin, J. R. (1985). Process and text: two aspects of human semiosis. In J. D. Benson & W. S. Greaves (Eds.) *Systemic perspectives on discourse, Volume 1; Selected theoretical papers from the ninth International Systemic Workshop*, (pp. 248–274). Norwood, New Jersey: Ablex Publishing Corporation.

Martin, J. R. (1992a). *English Text: System and Structure*. Amsterdam: John Benjamins Publishing.

Martin, J. R. (1992b). Write it right. Stage 1: Scientific literacy. Literacy in industry research project, NSW Department of School Education, Metropolitan East Disadvantaged Schools' Program, Sydney, Australia.

Martin, J. R. (1993). Genre and Literacy - Modeling Context in Educational Linguistics. *Annual Review of Applied Linguistics*, *13*(1), 141–172.

Martin, J. R. (1997). Analysing genre: functional parameters. In F. Christie & J. R. Martin (Eds.) *Genre and Institutions. Social Processes in the Workplace and School*, (pp. 3–39). London and Washington: Cassell.

Martin, J. R. (2007). Construing knowledge: a functional linguistic perspective. In F. Christie & J. R. Martin (Eds.) *Language, Knowledge and Pedagogy. Functional Linguistic and Sociological Perspectives*, (pp. 34–64). London, New York: Continuum.

Martin, J. R. (2009). Discourse studies. In M. A. K. Halliday & J. Webster (Eds.) *Continuum Companion to Systemic Functional Linguistic*, (pp. 154–165). London: Continuum.

Martin, J. R. & Rose, D. (2007). *Working with Discourse. Meaning beyond the clause*. London, New York: Continuum, 2. ed.

Martin, J. R. & Veel, R. (Eds.) (1998). *Reading science. Critical and functional perspectives on discourses of science*. London: Routledge.

Martín-Martín, P. (2003). A Genre Analysis of English and Spanish Research Paper Abstracts in Experimental Social Sciences. *English for Specific Purposes*, *22*(1), 25–43.

Martín-Martín, P. (2005). *The Rhetoric of the Abstract in English and Spanish Scientific Discourse. A Cross-Cultural Genre-Analytic Approach*, vol. 279 of *European University Studies*. Bern: Peter Lang.

Matthews, P. H. (1981). *Syntax*. Cambridge: Cambridge University Press.

Matthiessen, C. M. I. M. (1993). Register in the round: diversity in a unified theory of register analysis. In M. Ghadessy (Ed.) *Register Analysis. Theory and Practice*, (pp. 221–285). London and New York: Pinter Publishers.

Matthiessen, C. M. I. M. (2006). Frequency profiles of some basic grammatical systems: an interim report. In G. Thompson & S. Hunston (Eds.) *System and Corpus: Exploring connections*, (pp. 103–142). London: Equinox.

Matthiessen, C. M. I. M. (2007). The 'architecture' of language according to systemic functional theory: developments since the 1970s. In R. Hasan, C. Matthiessen & J. Webster (Eds.) *Continuing Discourse on Language. A functional perspective*, (pp. 505–561). London, Oakville: Equinox.

McEnery, T. & Wilson, A. (2001). *Corpus Linguistics: An Introduction*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh Univ Pr: Edinburgh University Press, 2nd ed.

Meadows, A. J. (Ed.) (1980). *Development of Science Publishing in Europe*. Amsterdam, New York, Oxford: Elsevier Science Publisher.

Moessner, L. (2009). How representative are the 'Philosophical Transactions of the Royal Society' of the 17th-century scientific writing? In A. Renouf & A. Kehoe (Eds.) *Corpus Linguistics: Refinements and Reassessments*, (pp. 221–238). Amsterdam, New York: Rodopi.

Montgomery, S. L. (1996). *The Scientific Voice*. New York, London: The Guilford Press.

Neumann, S. (2003). *Textsorten und Übersetzen*. Frankfurt am Main: Peter Lang.

Neumann, S. (2008). *Contrastive Register Variation. A quantitative approach to the comparison of English and German*. Unpublished. Habilitationsschrift.

Nwogu, K. N. (1991). Structure of science popularizations: A genre-analysis approach to the schema of popularized medical texts. *English for Specific Purposes*, *10*(2), 111–123.

Nwogu, K. N. (1993). Thematic Progression and Paragraph Development in the experimental Research Paper: implications for academic writing. *Rewiew of applied linguistics*, *101-102*, 89–104.

Nwogu, K. N. (1997). The medical research paper: Structure and functions. *English for Specific Purposes*, *16*(2), 119–138.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

OED Online (1989). The Oxford English Dictionary.
URL `http://dictionary.oed.com/`

Oesterreicher, W. (2001). Historizität - Sprachvariation, Sprachverschiedenheit, Sprachwandel. In M. Haspelmath, E. König, W. Oesterreicher & W. Raible (Eds.) *Language Typology and Language Universals / La typologie des langues et les universaux linguistiques / Sprachtypologie und sprachliche Universalien*, vol. 20.2, (pp. 1554–1595). Berlin, New York: Walter de Gruyter.

O'Halloran, K. (2005). *Mathematical Discourse: Language, Symbolism and Visual Images*. London [u.a.]: Continuum.

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

Pike, K. (1967). *Language in Relation to a Unified Theory of Structure of Human Behaviour*. The Hague: Mouton.

Randaccio, M. (2004). Language change in scientific discourse. *Journal of Science Communication (JCOM)*, *3*(2), 1–15.

Rasinger, S. M. (2008). *Quantitative Research in Linguistics. An Introduction*. Research Methods in Linguistics. London and New York: Continuum.

Reid, T. B. W. (1956). Linguistics, structuralism and philology. *Archivum Linguisticum VIII.*, *8*.

Saki, M. (2004). Thematic progression patterns and generic identity in abstracts. In D. Banks (Ed.) *Text and Texture*, (pp. 429–439). Paris, Budapest, Torino: L'Harmattan.

Salager-Meyer, F. (1990). Discoursal Movements in Medical English Abstracts and Their Linguistic Exponents: A Genre Analysis Study. *Interface: Journal of Applied Linguistics/Tijdschrift voor Toegepaste Lingüstiek (Interface)*, *4*(2), 107–124.

Salager-Meyer, F. (1992). A text-type and move analysis study of verb tense and modality distribution in medical English abstracts. *English for Specific Purposes*, *11*, 93–113.

Salager-Meyer, F. (1994). Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes*, *13*(2), 149–170.

Salager-Meyer, F. (1999). Referential behaviour in scientific writing: a diachronic study (1810-1995). *English for Specific Purposes*, *18*(3), 279–305.

Saville-Troike, M. (1982). *The ethongraphy of communication.* Oxford: Basil Blackwell.

Schmid, H. (1994a). Part-of-speech tagging with neural networks. In *In Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*.

Schmid, H. (1994b). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Scott, M. (2001). Mapping key words to *problem* and *solution.* In M. Scott & G. Thompson (Eds.) *Patterns of Text*, (pp. 109–127). John Benjamins.

Scott, M. (2008). *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. (1992). Trust the text. In M. Davies & L. Ravelli (Eds.) *Advances in Systemic Linguistics. Recent Theory and Practice*. London: Pinter Publishers.

Sinclair, J. (1996). The search for units of meaning. *Textus*, *IX*(1), 75–106.

Sinclair, J. (2003). Corpora for lexicography. In P. van Sterkenburg (Ed.) *A practical guide to lexicography*, (pp. 167–178). Amsterdam: Benjamins.

Steiner, E. (1983). *Die Entwicklung des Britischen Kontextualismus*. Heidelberg: Julius Groos Verlag.

Steiner, E. (1996). An extended register analysis as a form of text analysis for translation. In G. Wotjak & H. Schmidt (Eds.) *Modelle der Translation – Models of Translation*, (pp. 235–256). Leipzig: Leipziger Schriften zur Kultur-, Literatur-, Sprach- und Übersetzungswissenschaft.

Steiner, E., Hansen-Schirra, S. & Neumann, S. (2007). Cohesive explicitness and explicitation in an English-German translation corpus. *Languages in Contrast*, *7*(2), 241–266.

Stotesburry, H. (2003). Evaluation in research article abstracts in the narrative and hard sciences. *Journal of English for Academic Purposes*, *2*(4), 327–341.

Stubbs, M. (1986). Lexical density: A technique and some findings. In M. Coulthard (Ed.) *Talking About Text: Studies presented to David Brazil on his retirement, Discourse Analysis Monograph no. 13*, (pp. 27–42). Birmingham: English Language Research, Univ. of Birmingham.

Swales, J. M. (1981). Aspects of article introductions. Aston-ESP-research-reports NO. 1, The Language Studies Unit, The University of Aston, Birmingham, England. Sixth Impression - June 1987.

Swales, J. M. (1990). *Genre Analysis. English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. M. (2004). *Research Genres. Exploration and Applications*. Cambridge: Cambridge University Press.

Swales, J. M. & Feak, C. B. (2009). *Abstracts and the Writing of Abstracts (The Michigan Series in English for Academic & Professional Purposes)*. Ann Arbor: The University of Michigan Press.

Teich, E. (2003). *Cross-Linguistic Variation in System und Text. A Methodology for the Investigation of Translations and Comparable Texts*. Berlin: Mouton de Gruyter.

Teich, E. & Fankhauser, P. (2010). Exploring a corpus of scientific texts using data mining. In S. T. Gries, M. Davies & S. Wulff (Eds.) *Selected Papers from the American Conference on Corpus Linguistics (AACL) 2008, Provo, Utah*. Amsterdam: Rodopi.

Teich, E. & Holtz, M. (2009). Scientific registers in contact: An exploration of the lexico-grammatical properties of interdisciplinary discourses. *International Journal of Corpus Linguistics*, *14*(4), 524–548.

Teubert, W. (2010a). My brave old world. *International Journal of Corpus Linguistics*, *15*(3), 395–399.

Teubert, W. (2010b). Our brave new world? *International Journal of Corpus Linguistics*, *15*(3), 354–358.

Thompson, G. (2004). *Introducing Functional Grammar*. London: Arnold, 2nd ed.

Thompson, G. & Hunston, S. (Eds.) (2006). *System and Corpus: Exploring connections*. London: Equinox.

Threadgold, T. (1989). Talking about Genre : Ideologies and Incompatible Discourses. *Journal of Cultural Studies*, *3*(3), 101–127.

Tucker, G. (2006). Systemic in*corpora*tion: on the relationship between corpus and systemic functional grammar. In G. Thompson & S. Hunston (Eds.) *System and Corpus: Exploring connections*, chap. 5, (pp. 81–102). London: Equinox.

Ure, J. (1971). Lexical density and register differentiation. In G. E. Perren & J. L. M. Trim (Eds.) *Applications of Linguistics. Selected papers of the Second International Congress of Applied Linguistics, Cambridge 1969*, (pp. 443–452). Cambridge University Press.

Ure, J. (1982). Introduction: Approches to the study of register genre. *International Journal of the Sociology of Language: IJSL*, *35*, 5–23.

Ure, J. N. (1969a). Practical registers. *ELT Journal*, *XXIII*(2), 107–114.

Ure, J. N. (1969b). Practical registers. *ELT Journal*, *XXIII*(3), 206–215.

van Dijk, T. A. (1980). *Macrostructures*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Ventola, E. (1992). Writing scientific English: overcoming intercultural problems. *International Journal of Applied Linguistics*, *2*(2), 191–220.

Ventola, E. (1994). From Syntax to Text: Problems in Producing Scientific Abstracts in L_2. In S. Čmejrková & F. Štícha (Eds.) *The Syntax of Sentence and Text*. Amsterdam, Philadelphia: John Benjamins Publishing.

Ventola, E. (1996). Packing and unpacking of information in academic texts. In E. Ventola & A. Mauranen (Eds.) *Academic Writing. Intercultural and Textual Issues*, Pragmatics & Beyond: New Series (P&B): 41, (pp. 153–194). Amsterdam/Philadelphia: John Benjamins Publishing.

Ventola, E. (1997). Abstracts as an object of linguistic study. In F. Danes, E. Havlova & S. Cmejrkova (Eds.) *Writing vs. Speaking: Language, Text, Discourse, Communication. Proceedings of the Conference held at the Czech Language*, (pp. 333–352). Tübingen: Günter Narr.

Webster, J. J. (Ed.) (2009). *The Essential Halliday*. London and New York: Continuum.

Wignell, P. (1998). Technicality and abstraction in social science. In J. R. Martin & R. Veel (Eds.) *Reading science. Critical and functional perspectives on discourses of science*, (pp. 297–326). London [u.a.]: Routledge.

Wignell, P. (2007). Vertical and horizontal discourse and the social sciences. In F. Christie & J. R. Martin (Eds.) *Language, Knowledge and Pedagogy. Functional Linguistic and Sociological Perspectives*, (pp. 184–204). London, New York: Continuum.

Wignell, P., Martin, J. & Eggins, S. (1993). The discourse of geography: Ordering and explaning the experiential world. In M. A. K. Halliday & J. Martin (Eds.) *Writing Science: Literacy and Discursive Power*, (pp. 136–165). London: University of Pittsburgh Press.

Appendices

A.1 Example of multi-layer annotation

This is the result of the multi-layer annotation in XML generated by AnnoLab after tokenization, part-of-speech tagging, lemmatization, and syntactic parsing of the sentence: `This is an example.`

```
<?xml version="1.0" encoding="UTF-8"?>
<gam:root xmlns:gam="http://www.linglit.tu-darmstadt.de/PACE/GAM">
  <gam:headers>
    <gam:header gam:id="3451" gam:name="example-annolab"/>
    <gam:header gam:id="3501" gam:name="Parse"/>
    <gam:header gam:id="3503" gam:name="Sentence"/>
    <gam:header gam:id="3502" gam:name="Token"/>
  </gam:headers>
  <gam:annotations>
    <gam:layer gam:id="3501" gam:name="Parse">
      <ROOT>
        <Constituent cat="ROOT">
          <Constituent cat="S">
            <Constituent cat="NP">
              <Constituent cat="DT">
                <Token posTag="DT">
                  <gam:a gam:id="11"/>
                </Token>
              </Constituent>
            </Constituent>
            <Constituent cat="VP">
              <Constituent cat="VBZ">
                <Token posTag="VBZ">
                  <gam:a gam:id="12"/>
                </Token>
              </Constituent>
              <Constituent cat="NP">
                <Constituent cat="DT">
                  <Token posTag="DT">
                    <gam:a gam:id="13"/>
                  </Token>
                </Constituent>
                <Constituent cat="NN">
                  <Token posTag="NN">
                    <gam:a gam:id="14"/>
                  </Token>
                </Constituent>
              </Constituent>
            </Constituent>
            <Constituent cat=".">
              <Token posTag=".">
                <gam:a gam:id="15"/>
              </Token>
            </Constituent>
          </Constituent>
        </Constituent>
      </ROOT>
    </gam:layer>
    <gam:layer gam:id="3503" gam:name="Sentence">
      <segments>
        <segment>
          <gam:a gam:id="0"/>
        </segment>
      </segments>
    </gam:layer>
    <gam:layer gam:id="3502" gam:name="Token">
      <segments>
        <segment posTag="DT">
          <gam:a gam:id="1"/>
        </segment>
        <segment>
          <gam:a gam:id="2"/>
        </segment>
        <segment posTag="VBZ">
          <gam:a gam:id="3"/>
        </segment>
        <segment>
          <gam:a gam:id="4"/>
        </segment>
        <segment posTag="DT">
          <gam:a gam:id="5"/>
        </segment>
        <segment>
          <gam:a gam:id="6"/>
        </segment>
        <segment posTag="NN">
          <gam:a gam:id="7"/>
        </segment>
        <segment>
          <gam:a gam:id="8"/>
        </segment>
        <segment posTag=".">
          <gam:a gam:id="9"/>
        </segment>
        <segment>
          <gam:a gam:id="10"/>
        </segment>
      </segments>
    </gam:layer>
  </gam:annotations>
  <gam:layout gam:sig="annolab://default/example-annolab">
    <gam:root>
      <gam:seg gam:type="seq" gam:sig="default:3451" gam:s="0" gam:e="4">
```

```
        <gam:content>This</gam:content>
        <gam:ref gam:aid="0"/>
        <gam:ref gam:aid="1"/>
        <gam:ref gam:aid="2"/>
        <gam:ref gam:aid="11"/>
      </gam:seg>
      <gam:seg gam:type="seq" gam:sig="default:3451" gam:s="4" gam:e="5">
        <gam:content> </gam:content>
        <gam:ref gam:aid="0"/>
      </gam:seg>
      <gam:seg gam:type="seq" gam:sig="default:3451" gam:s="5" gam:e="7">
        <gam:content>is</gam:content>
        <gam:ref gam:aid="0"/>
        <gam:ref gam:aid="3"/>
        <gam:ref gam:aid="4"/>
        <gam:ref gam:aid="12"/>
      </gam:seg>
      <gam:seg gam:type="seq" gam:sig="default:3451" gam:s="7" gam:e="8">
        <gam:content> </gam:content>
        <gam:ref gam:aid="0"/>
      </gam:seg>
      <gam:seg gam:type="seq" gam:sig="default:3451" gam:s="8" gam:e="10">
        <gam:content>an</gam:content>
        <gam:ref gam:aid="0"/>
        <gam:ref gam:aid="5"/>
        <gam:ref gam:aid="6"/>
        <gam:ref gam:aid="13"/>
      </gam:seg>
      <gam:seg gam:type="seq" gam:sig="default:3451" gam:s="10" gam:e="11">
        <gam:content> </gam:content>
        <gam:ref gam:aid="0"/>
      </gam:seg>
      <gam:seg gam:type="seq" gam:sig="default:3451" gam:s="11" gam:e="18">
        <gam:content>example</gam:content>
        <gam:ref gam:aid="0"/>
        <gam:ref gam:aid="7"/>
        <gam:ref gam:aid="8"/>
        <gam:ref gam:aid="14"/>
      </gam:seg>
      <gam:seg gam:type="seq" gam:sig="default:3451" gam:s="18" gam:e="19">
        <gam:content>.</gam:content>
        <gam:ref gam:aid="0"/>
        <gam:ref gam:aid="9"/>
        <gam:ref gam:aid="10"/>
        <gam:ref gam:aid="15"/>
      </gam:seg>
      <gam:seg gam:type="seq" gam:sig="default:3451" gam:s="19" gam:e="20">
        <gam:content>
</gam:content>
      </gam:seg>
      <gam:seg gam:type="seq" gam:sig="default:3451" gam:s="0" gam:e="20"/>
    </gam:root>
  </gam:layout>
</gam:root>
```

A.2 List of tags with corresponding part-of-speech

List of tags with corresponding part-of-speech of the Penn Treebank[50] tagset.

Tag	Part-of-speech
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective comparative
JJS	Adjective superlative
LS	List item marker
MD	Modal
NN	Noun singular or mass
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
SYM	Symbol
TO	to
UH	Interjection
VB	Verb base form
VBD	Verb past tense
VBG	Verb gerund or present participle
VBN	Verb past participle
VBP	Verb, non-3rd person singular present
VVN	Verb, past participle
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

[50] URL: `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/` (accessed: 31 July 2010).

A.3 Distribution of lexical words in the AbstRA corpus

Table A.1 and Table A.2 show the distribution of lexical words for abstracts and RAs in each discipline, respectively. The values of the frequency of occurrence of lexical words is given both as raw frequency and as relative frequency, i.e., raw frequency divided by the total number of tokens for each single text.

Adjectives		Nouns		Personal Pronouns		Adverbs		Verbs	
F	RF	F	RF	F	RF	F	RF	F	RF
Computer science									
3	0.0857	12	0.3429	1	0.0286	0	0.0000	3	0.0857
11	0.1209	17	0.1868	3	0.0330	1	0.0110	18	0.1978
15	0.0811	47	0.2541	5	0.0270	1	0.0054	15	0.0811
23	0.1018	56	0.2478	3	0.0133	11	0.0487	29	0.1283
9	0.0818	33	0.3000	2	0.0182	1	0.0091	17	0.1545
15	0.1042	48	0.3333	4	0.0278	0	0.0000	17	0.1181
21	0.1567	39	0.2910	0	0.0000	3	0.0224	17	0.1269
6	0.0472	46	0.3622	0	0.0000	4	0.0315	21	0.1654
12	0.0952	38	0.3016	2	0.0159	2	0.0159	16	0.1270
12	0.0833	34	0.2361	4	0.0278	2	0.0139	16	0.1111
19	0.0411	147	0.3182	5	0.0108	7	0.0152	37	0.0801
7	0.0729	28	0.2917	2	0.0208	0	0.0000	7	0.0729
9	0.0750	31	0.2583	2	0.0167	1	0.0083	9	0.0750
16	0.0800	50	0.2500	8	0.0400	5	0.0250	34	0.1700
10	0.0935	29	0.2710	0	0.0000	4	0.0374	16	0.1495
5	0.0676	25	0.3378	4	0.0541	1	0.0135	9	0.1216
27	0.1579	42	0.2456	1	0.0058	7	0.0409	38	0.2222
Linguistics									
5	0.0227	68	0.3091	5	0.0227	3	0.0136	31	0.1409
13	0.0414	100	0.3185	1	0.0032	7	0.0223	9	0.0287
20	0.0870	57	0.2478	7	0.0304	11	0.0478	34	0.1478
27	0.1184	65	0.2851	3	0.0132	5	0.0219	30	0.1316
9	0.0448	66	0.3284	1	0.0050	5	0.0249	19	0.0945
16	0.1081	40	0.2703	2	0.0135	4	0.0270	22	0.1486
11	0.0866	34	0.2677	2	0.0157	5	0.0394	11	0.0866
22	0.0712	64	0.2071	5	0.0162	9	0.0291	31	0.1003
8	0.0462	57	0.3295	1	0.0058	6	0.0347	18	0.1040
27	0.1000	81	0.3000	2	0.0074	11	0.0407	39	0.1444
22	0.0824	82	0.3071	1	0.0037	7	0.0262	25	0.0936
21	0.0913	68	0.2957	5	0.0217	7	0.0304	29	0.1261
14	0.1000	37	0.2643	4	0.0286	4	0.0286	15	0.1071
22	0.0887	71	0.2863	4	0.0161	6	0.0242	34	0.1371
20	0.0873	57	0.2489	3	0.0131	7	0.0306	42	0.1834
15	0.1948	17	0.2208	1	0.0130	0	0.0000	9	0.1169
8	0.0889	27	0.3000	0	0.0000	2	0.0222	8	0.0889
28	0.1228	54	0.2368	1	0.0044	9	0.0395	28	0.1228
13	0.1083	37	0.3083	2	0.0167	3	0.0250	16	0.1333
23	0.1237	62	0.3333	3	0.0161	4	0.0215	17	0.0914
17	0.0787	49	0.2269	6	0.0278	8	0.0370	33	0.1528

Table A.1 – Continued on next page

F = raw frequency; RF = relative frequency

Table A.1 – continued from previous page

Adjectives		Nouns		Personal Pronouns		Adverbs		Verbs	
F	RF	F	RF	F	RF	F	RF	F	RF
27	0.1063	69	0.2717	5	0.0197	17	0.0669	29	0.1142
20	0.1149	51	0.2931	0	0.0000	10	0.0575	18	0.1034
13	0.1226	25	0.2358	2	0.0189	3	0.0283	14	0.1321
Biology									
10	0.0376	74	0.2782	2	0.0075	5	0.0188	24	0.0902
21	0.0843	82	0.3293	3	0.0120	11	0.0442	22	0.0884
16	0.0812	63	0.3198	4	0.0203	3	0.0152	31	0.1574
24	0.1558	42	0.2727	1	0.0065	6	0.0390	20	0.1299
31	0.0917	116	0.3432	1	0.0030	4	0.0118	28	0.0828
24	0.0752	92	0.2884	2	0.0063	5	0.0157	28	0.0878
15	0.1034	39	0.2690	3	0.0207	4	0.0276	22	0.1517
22	0.1073	65	0.3171	3	0.0146	5	0.0244	30	0.1463
20	0.1212	63	0.3818	0	0.0000	5	0.0303	18	0.1091
17	0.1197	39	0.2746	2	0.0141	3	0.0211	22	0.1549
23	0.0895	90	0.3502	2	0.0078	6	0.0233	27	0.1051
15	0.0781	62	0.3229	3	0.0156	3	0.0156	25	0.1302
32	0.1081	98	0.3311	1	0.0034	8	0.0270	32	0.1081
23	0.0839	93	0.3394	5	0.0182	11	0.0401	33	0.1204
5	0.0485	28	0.2718	1	0.0097	11	0.1068	16	0.1553
18	0.1023	65	0.3693	0	0.0000	3	0.0170	22	0.1250
18	0.1304	52	0.3768	1	0.0072	3	0.0217	14	0.1014
26	0.1398	57	0.3065	3	0.0161	5	0.0269	30	0.1613
11	0.0671	58	0.3537	1	0.0061	2	0.0122	23	0.1402
9	0.0577	49	0.3141	2	0.0128	7	0.0449	24	0.1538
43	0.1792	63	0.2625	1	0.0042	11	0.0458	34	0.1417
13	0.0839	59	0.3806	2	0.0129	5	0.0323	16	0.1032
16	0.0576	95	0.3417	1	0.0036	5	0.0180	34	0.1223
Mechanical engineering									
8	0.0615	44	0.3385	0	0.0000	0	0.0000	17	0.1308
24	0.1057	80	0.3524	2	0.0088	4	0.0176	25	0.1101
12	0.1101	40	0.3670	0	0.0000	2	0.0183	12	0.1101
15	0.0789	61	0.3211	0	0.0000	6	0.0316	18	0.0947
9	0.0441	72	0.3529	1	0.0049	5	0.0245	27	0.1324
13	0.0807	46	0.2857	0	0.0000	4	0.0248	26	0.1615
9	0.0545	50	0.3030	3	0.0182	4	0.0242	24	0.1455
11	0.0853	44	0.3411	2	0.0155	3	0.0233	20	0.1550
20	0.1370	44	0.3014	0	0.0000	9	0.0616	19	0.1301
16	0.0860	65	0.3495	0	0.0000	10	0.0538	22	0.1183
14	0.0645	67	0.3088	0	0.0000	2	0.0092	23	0.1060
14	0.0791	55	0.3107	2	0.0113	5	0.0282	28	0.1582
5	0.0500	40	0.4000	1	0.0100	2	0.0200	14	0.1400
5	0.0431	46	0.3966	0	0.0000	1	0.0086	16	0.1379
6	0.0909	18	0.2727	1	0.0152	1	0.0152	12	0.1818
8	0.0620	44	0.3411	0	0.0000	3	0.0233	19	0.1473
16	0.1103	55	0.3793	0	0.0000	2	0.0138	19	0.1310
26	0.0939	94	0.3394	0	0.0000	2	0.0072	26	0.0939
9	0.0769	35	0.2991	0	0.0000	3	0.0256	13	0.1111
11	0.0815	51	0.3778	0	0.0000	0	0.0000	10	0.0741
14	0.0741	62	0.3280	3	0.0159	3	0.0159	29	0.1534
21	0.1909	27	0.2455	1	0.0091	5	0.0455	13	0.1182
12	0.1053	33	0.2895	0	0.0000	2	0.0175	6	0.0526
14	0.1261	41	0.3694	0	0.0000	2	0.0180	14	0.1261
9	0.0909	27	0.2727	0	0.0000	0	0.0000	19	0.1919

Table A.1 – Continued on next page

F = raw frequency; RF = relative frequency

Table A.1 – continued from previous page

Adjectives		Nouns		Personal Pronouns		Adverbs		Verbs	
F	RF	F	RF	F	RF	F	RF	F	RF
9	0.0726	43	0.3468	2	0.0161	5	0.0403	15	0.1210
25	0.1157	69	0.3194	1	0.0046	3	0.0139	22	0.1019
19	0.1226	42	0.2710	1	0.0065	2	0.0129	24	0.1548
16	0.1127	41	0.2887	0	0.0000	1	0.0070	13	0.0915

Table A.1: Distribution of lexical words for abstracts in the AbstRA corpus

Adjectives		Nouns		Personal Pronouns		Adverbs		Verbs	
F	RF	F	RF	F	RF	F	RF	F	RF
Computer science									
155	0.0654	609	0.2571	33	0.0139	89	0.0376	343	0.1448
371	0.0812	943	0.2063	92	0.0201	201	0.0440	675	0.1477
359	0.0657	1310	0.2398	135	0.0247	244	0.0447	707	0.1294
372	0.0639	1571	0.2697	101	0.0173	184	0.0316	732	0.1257
505	0.0816	1488	0.2403	100	0.0162	177	0.0286	756	0.1221
246	0.0659	1107	0.2966	79	0.0212	91	0.0244	428	0.1147
343	0.0763	1160	0.2582	79	0.0176	145	0.0323	576	0.1282
422	0.0701	1721	0.2858	82	0.0136	202	0.0335	880	0.1462
220	0.0647	794	0.2335	81	0.0238	132	0.0388	460	0.1353
238	0.0555	992	0.2315	92	0.0215	199	0.0464	520	0.1213
271	0.0515	1488	0.2827	77	0.0146	149	0.0283	610	0.1159
417	0.0740	1368	0.2427	118	0.0209	184	0.0326	693	0.1229
100	0.0795	330	0.2623	10	0.0079	34	0.0270	179	0.1423
1248	0.0809	4075	0.2641	374	0.0242	541	0.0351	1947	0.1262
191	0.0843	551	0.2433	44	0.0194	83	0.0366	290	0.1280
137	0.0597	599	0.2610	48	0.0209	92	0.0401	301	0.1312
341	0.1136	660	0.2199	37	0.0123	106	0.0353	374	0.1246
350	0.0655	1434	0.2685	132	0.0247	173	0.0324	682	0.1277
257	0.0564	1229	0.2696	175	0.0384	169	0.0371	726	0.1592
838	0.0945	2240	0.2526	184	0.0208	323	0.0364	1140	0.1286
377	0.0784	1380	0.2871	83	0.0173	159	0.0331	588	0.1223
181	0.0733	662	0.2682	29	0.0118	91	0.0369	312	0.1264
303	0.0650	1192	0.2556	37	0.0079	125	0.0268	537	0.1151
180	0.0562	723	0.2257	52	0.0162	151	0.0471	469	0.1464
205	0.0641	712	0.2226	86	0.0269	185	0.0578	504	0.1576
259	0.0558	1266	0.2729	120	0.0259	141	0.0304	589	0.1270
752	0.0646	3213	0.2759	185	0.0159	458	0.0393	1477	0.1268
Linguistics									
592	0.0749	2050	0.2595	81	0.0103	269	0.0341	968	0.1225
948	0.0870	2973	0.2727	171	0.0157	422	0.0387	1228	0.1127
1239	0.1017	2866	0.2353	192	0.0158	631	0.0518	1507	0.1237
984	0.0785	3044	0.2430	179	0.0143	529	0.0422	1747	0.1394
1308	0.0846	3376	0.2183	378	0.0244	829	0.0536	2059	0.1331
700	0.0990	1705	0.2410	196	0.0277	211	0.0298	909	0.1285
359	0.0722	1291	0.2597	63	0.0127	199	0.0400	577	0.1160
754	0.0766	2537	0.2576	155	0.0157	365	0.0371	1245	0.1264

Table A.2 – Continued on next page

F = raw frequency; RF = relative frequency

Table A.2 – continued from previous page

Adjectives		Nouns		Personal		Adverbs		Verbs	
				Pronouns					
F	RF	F	RF	F	RF	F	RF	F	RF
376	0.0624	1534	0.2546	75	0.0124	192	0.0319	738	0.1225
1028	0.0902	3116	0.2734	100	0.0088	367	0.0322	1337	0.1173
669	0.0786	1904	0.2237	106	0.0125	360	0.0423	1088	0.1278
1079	0.0919	2835	0.2414	213	0.0181	520	0.0443	1249	0.1063
134	0.0866	369	0.2384	16	0.0103	63	0.0407	172	0.1111
477	0.0751	1457	0.2294	99	0.0156	252	0.0397	933	0.1469
Biology									
227	0.0605	1099	0.2931	17	0.0045	104	0.0277	383	0.1021
251	0.0871	817	0.2836	27	0.0094	119	0.0413	302	0.1048
237	0.0522	1445	0.3181	32	0.0070	121	0.0266	557	0.1226
136	0.0688	653	0.3305	4	0.0020	40	0.0202	194	0.0982
187	0.0696	832	0.3096	31	0.0115	66	0.0246	303	0.1128
306	0.0754	1399	0.3448	7	0.0017	67	0.0165	327	0.0806
213	0.0645	920	0.2786	19	0.0058	114	0.0345	315	0.0954
286	0.0864	928	0.2804	33	0.0100	137	0.0414	353	0.1067
180	0.0653	866	0.3143	8	0.0029	72	0.0261	358	0.1299
162	0.0671	816	0.3382	8	0.0033	68	0.0282	285	0.1181
228	0.0767	1000	0.3364	16	0.0054	72	0.0242	320	0.1076
178	0.0697	878	0.3440	11	0.0043	39	0.0153	267	0.1046
252	0.0569	1395	0.3148	37	0.0084	135	0.0305	591	0.1334
352	0.0826	1286	0.3016	17	0.0040	105	0.0246	489	0.1147
328	0.0731	1304	0.2907	43	0.0096	156	0.0348	595	0.1327
252	0.0712	1076	0.3039	24	0.0068	113	0.0319	433	0.1223
181	0.0739	872	0.3561	10	0.0041	69	0.0282	301	0.1229
210	0.0788	944	0.3542	6	0.0023	54	0.0203	217	0.0814
300	0.0828	1117	0.3082	21	0.0058	87	0.0240	404	0.1115
270	0.0701	1211	0.3143	16	0.0042	75	0.0195	426	0.1106
256	0.0821	966	0.3097	36	0.0115	132	0.0423	356	0.1141
296	0.1251	693	0.2929	19	0.0080	89	0.0376	253	0.1069
167	0.0607	958	0.3484	35	0.0127	77	0.0280	331	0.1204
269	0.0485	1667	0.3004	28	0.0050	106	0.0191	599	0.1079
Mechanical engineering									
202	0.0688	809	0.2755	25	0.0085	48	0.0163	300	0.1022
307	0.0743	1306	0.3160	27	0.0065	97	0.0235	439	0.1062
167	0.0675	814	0.3289	5	0.0020	55	0.0222	273	0.1103
221	0.0955	690	0.2983	10	0.0043	52	0.0225	264	0.1141
159	0.0696	783	0.3430	10	0.0044	63	0.0276	234	0.1025
271	0.0771	1008	0.2866	12	0.0034	104	0.0296	495	0.1407
108	0.0601	504	0.2805	15	0.0083	51	0.0284	241	0.1341
222	0.0636	1061	0.3039	25	0.0072	108	0.0309	420	0.1203
190	0.0873	642	0.2949	7	0.0032	64	0.0294	283	0.1300
268	0.0777	1031	0.2991	9	0.0026	122	0.0354	384	0.1114
183	0.0716	764	0.2990	6	0.0023	41	0.0160	343	0.1342
233	0.0923	676	0.2678	16	0.0063	79	0.0313	341	0.1351
197	0.0723	771	0.2830	18	0.0066	65	0.0239	364	0.1336
159	0.0725	547	0.2494	19	0.0087	61	0.0278	254	0.1158
216	0.0850	734	0.2889	16	0.0063	49	0.0193	328	0.1291
285	0.0790	1105	0.3063	12	0.0033	114	0.0316	401	0.1111
271	0.0820	1003	0.3036	16	0.0048	74	0.0224	303	0.0917
178	0.0906	609	0.3099	10	0.0051	32	0.0163	184	0.0936
132	0.0693	479	0.2514	23	0.0121	36	0.0189	212	0.1113
237	0.0793	819	0.2739	19	0.0064	92	0.0308	327	0.1094
165	0.0727	685	0.3016	11	0.0048	59	0.0260	311	0.1369

Table A.2 – Continued on next page

F = raw frequency; RF = relative frequency

Table A.2 – continued from previous page

Adjectives		Nouns		Personal Pronouns		Adverbs		Verbs	
F	RF	F	RF	F	RF	F	RF	F	RF
267	0.1207	488	0.2206	33	0.0149	93	0.0420	265	0.1198
196	0.0790	588	0.2369	43	0.0173	80	0.0322	263	0.1060
122	0.0733	524	0.3149	7	0.0042	32	0.0192	155	0.0931
167	0.0763	616	0.2814	4	0.0018	44	0.0201	254	0.1160
191	0.0604	961	0.3039	23	0.0073	85	0.0269	363	0.1148
251	0.0844	822	0.2765	24	0.0081	80	0.0269	312	0.1049
272	0.0716	1161	0.3058	12	0.0032	77	0.0203	499	0.1314
257	0.0682	1119	0.2968	23	0.0061	64	0.0170	525	0.1393

Table A.2: Distribution of lexical words for RAs in the AbstRA corpus

A.4 Passive voice querying

Based on the work of Gustafsson (2006), the distribution of passive voice across the several disciplines of the corpus under study in this research was determined using IMS/CQP for PoS-based querying. The queries are:

- `passive-VB-A=[pos="VB"][]{0,3}[pos="VVN"&(uri=".*/A/.*")];`

  (e.g., *can be solved, to be solved, may be chosen*)
- `passive-VBZ-A=[pos="VBZ"][]{0,3}[pos="VVN"&(uri=".*/A/.*")];`

  (e.g., *is proved, is shown, is given*)
- `passive-VBP-not-VBG-A=[pos="VBP"]`
  `[!(pos="VBG")]{0,3}[pos="VVN"&(uri=".*/A/.*")];`

  (e.g., *are required, are in nature distributed, are further restricted*)
- `passive-VBD-A=[pos="VBD"][]{0,3}`
  `[pos="VVN"&(uri=".*/A/.*")];`

  (e.g., *were achieved, was achieved*)
- `passive-VBN-A=[pos="VBN"][]{0,3}`
  `[pos="VVN"&(uri=".*/A/.*")];`

  (e.g., *has/have been developed)*
- `passive-VBG-A=[pos="VBG"][]{0,3}`
  `[pos="VVN"&(uri=".*/A/.*")];`

  (e.g., *are being stopped*)
- `passive-VHZ-not-VBN-A=[pos="VHZ"][!(pos="VBN")]{0,3}`
  `[pos="VVN"&(uri=".*/A/.*")];`

  (e.g., *has been shown, has been proposed*)
- `passive-VHP-not-VBN-A=[pos="VHP"][!(pos="VBN")]{0,3}`
  `[pos="VVN"&(uri=".*/A/.*")];`

  (e.g., *systems of demonstratives have in general hitherto been treated as inherently spatial, have been proposed, have been develop, have been obtained*)
- `passive-shall-VB-VBN-A=[word="shall"][]{0,3}`
  `[pos="VB"][]{0,3}[pos="VVN"&(uri=".*/A/.*")];`

  (e.g., *that shall be used in future sections*)
- `passive-will-VB-VBN-A=[word="will"][]{0,3}`
  `[pos="VB"][]{0,3}[pos="VVN"&(uri=".*/A/.*")];`

  (e.g., *this textual unit will be presented together with a number*).

- `passive-has-VBN-A=[word="has"][pos="VBN"][]{0,3}`
  `[pos="VVN"&(uri=".*/A/.*")];`
- `passive-have-VBN-A=[word="have"][pos="VBN"][]{0,3}`
  `[pos="VVN"&(uri=".*/A/.*")];`
- `passive-had-VBN-C2=[word="had"][]{0,3}[pos="VBN"][]{0,3}`
  `[pos="VVN"&(uri=".*/C2/.*")];`
- `passive-modals-A=[pos="MD"&!(word="will")][]{0,3}`
  `[pos="VB"][]{0,3}[pos="VVN"&(uri=".*/A/.*")];`
  (e.g., *should be solved, may be chosen*)
- `passive-imperative-A=[(word="let")|(word="Let")][]{0,3}`
  `[word="be"][]{0,3}[pos="VVN"&(uri=".*/A/.*")];`
  (e.g., *Let S be the tableau obtained by rooting S*)

Where `(uri=".*/A/.*")` is as example of the selected discipline, computer science. It is replaced by C1, C2, and C3 for the disciplines of linguistics, biology, and mechanical engineering, respectively.

A.5 Data for inductive analysis

The complete data used for the inductive analysis is shown in Table A.3. This table is precisely equal to the matrix loaded in R for the hierarchical agglomerative cluster analysis and for the principal component analysis. Each of the rows represents a single text of the AbstRA corpus. Each column represent a vector (column variable) which specify the different features tested in this research. The columns in Table A.3 are numbered only but not named, due to space constrains for fitting all columns of this matrix in a single page. The column numbers represent the following features:
column 1: text
column 2: type
column 3: domain
column 4: tokens
column 5: prepositions
column 6: adjectives
column 7: modals
column 8: nouns
column 9: personal.pronouns
column 10: possessive.pronouns
column 11: adverbs
column 12: present.tense
column 13: past.participle
column 14: past.tense
column 15: nominalizations
column 16: sentence.length
column 17: lexical.density

The values in each row are either directly the raw value of frequency of occurrence of a given feature or a sum of several parts-of-speech indicating a given feature, as for example in the case of present and past tense, or the result of a formula, such as lexical density. The fact that the information concerning passives is not available for each single text (cf. Section 5.1.3.2) is partially solved by adopting the values of frequency of occurrence of participle forms. Although not every single occurrence of a particle implies a passive, each passive imply obligatorily the existence of a passive verb form.

The text *abstract.C2.5* is deleted before the inductive analysis is conducted in R since this text is misbuilt (cf. Section 5.1.1.2).

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
abstract.A.1	abstract	A	35	4	3	0	12	1	0	0	1	0	0	2	29	18
abstract.A.10	abstract	A	144	15	12	0	34	4	0	2	8	6	0	4	20.2	8.6
abstract.A.11	abstract	A	462	40	19	1	147	5	5	7	18	8	1	13	24.33	12.3
abstract.A.12	abstract	A	96	10	7	0	28	2	0	0	4	2	0	4	18.5	10.8
abstract.A.13	abstract	A	120	17	9	1	31	2	0	1	6	3	0	3	20.67	11
abstract.A.14	abstract	A	200	17	16	0	50	8	1	5	19	8	0	4	20.1	5.6
abstract.A.15	abstract	A	107	10	10	1	29	0	1	4	4	6	0	6	23.9	19.7
abstract.A.16	abstract	A	74	5	5	0	25	4	0	1	5	0	1	3	18.57	6.7
abstract.A.17	abstract	A	171	16	27	1	42	1	0	7	11	13	0	12	19.56	10.1
abstract.A.18	abstract	A	220	22	5	0	68	5	1	3	13	11	2	5	18.1	7.8
abstract.A.19	abstract	A	314	14	13	0	100	1	3	7	6	0	0	1	26	24.8
abstract.A.2	abstract	A	91	14	11	0	17	3	2	1	8	4	0	2	28.33	6.9
abstract.A.20	abstract	A	230	25	20	1	57	7	0	11	14	6	0	10	20.78	8.9
abstract.A.21	abstract	A	228	19	27	1	65	3	3	5	10	5	0	17	29.6	13.3
abstract.A.22	abstract	A	201	22	9	1	66	1	0	5	7	3	1	2	32.33	15.4
abstract.A.23	abstract	A	148	20	16	2	40	2	2	4	10	6	0	8	27.6	10.5
abstract.A.24	abstract	A	127	10	11	0	34	2	0	5	4	4	0	0	23.5	16
abstract.A.25	abstract	A	309	16	22	0	64	5	1	9	14	12	1	7	19.43	10.3
abstract.A.26	abstract	A	173	21	8	3	57	1	1	6	8	6	2	9	22.67	12.7
abstract.A.27	abstract	A	270	26	27	2	81	2	0	11	18	17	0	14	19.3	11.1
abstract.A.3	abstract	A	185	16	15	1	47	5	1	1	2	1	3	7	27.43	24
abstract.A.4	abstract	A	226	34	23	0	56	3	2	11	12	8	0	1	20.93	10.2
abstract.A.5	abstract	A	110	12	9	0	33	2	1	1	7	4	1	2	23.75	7.8
abstract.A.6	abstract	A	144	13	15	0	48	4	2	0	7	2	1	10	28.67	9.9
abstract.A.7	abstract	A	134	13	21	2	39	0	0	3	6	4	1	3	22.8	16
abstract.A.8	abstract	A	127	16	6	1	46	0	0	4	8	5	0	4	24.33	10.9
abstract.A.9	abstract	A	126	15	12	2	38	2	0	2	6	6	0	10	25.83	11.5
abstract.C1.1	abstract	C1	267	34	22	4	82	1	0	7	10	4	0	13	25.56	17.8
abstract.C1.10	abstract	C1	186	24	23	0	62	3	0	4	7	4	2	14	22.45	12.1
abstract.C1.11	abstract	C1	216	23	17	2	49	6	0	8	5	5	9	14	16.75	7.7
abstract.C1.12	abstract	C1	254	31	27	0	69	5	3	17	15	6	0	10	17.16	9.7
abstract.C1.13	abstract	C1	174	19	20	3	51	0	0	10	10	3	0	13	27.67	12.6
abstract.C1.14	abstract	C1	106	12	13	0	25	2	0	3	6	5	0	6	32	9.2
abstract.C1.2	abstract	C1	230	27	21	0	68	5	0	7	13	7	1	7	15.5	9.9
abstract.C1.3	abstract	C1	140	16	14	0	37	4	0	4	4	2	2	5	33	11.8
abstract.C1.4	abstract	C1	248	34	22	0	71	4	0	6	11	16	2	17	23.4	10.5
abstract.C1.5	abstract	C1	229	30	20	1	57	3	1	7	22	13	0	10	21.5	6.6

Table A.3 – Continued on next page

Table A.3 – continued from previous page

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
abstract.C1.6	abstract	C1	77	11	15	0	17	1	0	0	5	3	0	5	22	8.4
abstract.C1.7	abstract	C1	90	13	8	0	27	0	0	2	6	0	0	3	23.25	7.5
abstract.C1.8	abstract	C1	228	33	28	1	54	1	1	9	9	4	2	13	30.86	10.8
abstract.C1.9	abstract	C1	120	16	13	0	37	2	0	3	8	5	0	9	23.86	8.6
abstract.C2.1	abstract	C2	266	15	10	0	74	2	1	5	7	7	0	6	22.6	18.1
abstract.C2.10	abstract	C2	165	19	20	1	63	0	0	5	2	5	6	10	20.25	14.3
abstract.C2.11	abstract	C2	142	15	17	0	39	2	0	3	8	6	0	2	22.83	10.8
abstract.C2.12	abstract	C2	257	30	23	0	90	2	1	6	4	12	6	8	23	14.6
abstract.C2.13	abstract	C2	192	19	15	2	62	3	0	3	10	4	2	7	18.4	8.9
abstract.C2.14	abstract	C2	296	35	32	2	98	1	0	8	7	8	8	16	42.5	12.3
abstract.C2.15	abstract	C2	274	37	23	0	93	5	1	11	5	14	7	15	33	12.4
abstract.C2.16	abstract	C2	103	13	5	0	28	1	0	11	7	6	2	1	23.67	7.1
abstract.C2.17	abstract	C2	176	16	18	0	65	0	0	3	3	8	4	8	30	15.9
abstract.C2.18	abstract	C2	138	14	18	0	52	1	0	3	6	5	0	9	29.67	14.5
abstract.C2.19	abstract	C2	186	18	26	0	57	3	0	5	6	7	7	10	26.22	8.8
abstract.C2.2	abstract	C2	249	20	21	0	82	3	0	11	7	7	4	10	19.57	12
abstract.C2.20	abstract	C2	164	22	11	0	58	1	0	2	9	10	2	12	17.56	7.5
abstract.C2.21	abstract	C2	156	23	9	4	49	2	0	7	5	9	0	5	30.83	13
abstract.C2.22	abstract	C2	240	30	43	0	63	1	0	11	5	11	9	16	26	11
abstract.C2.23	abstract	C2	155	24	13	1	59	2	0	5	1	4	6	5	26.75	13.4
abstract.C2.24	abstract	C2	278	20	16	0	95	1	1	5	0	11	15	7	25.83	10.3
abstract.C2.3	abstract	C2	197	22	16	0	63	4	1	3	11	13	2	8	28	7.7
abstract.C2.4	abstract	C2	154	21	24	0	42	1	0	6	8	9	0	7	20.83	9.6
abstract.C2.5	abstract	C2	2633	30	14	5	1278	3	1	6	4	4	1	4	33	263.6
abstract.C2.6	abstract	C2	338	25	31	0	116	1	0	4	10	8	3	14	19.67	15.5
abstract.C2.7	abstract	C2	319	41	24	0	92	2	1	5	6	7	10	10	43.75	9.9
abstract.C2.8	abstract	C2	145	20	15	1	39	3	0	4	8	7	2	8	19.2	8.9
abstract.C2.9	abstract	C2	205	23	22	0	65	3	1	5	8	11	0	4	31	9.4
abstract.C3.1	abstract	C3	130	23	8	0	44	0	0	0	5	5	1	9	20.29	11.2
abstract.C3.10	abstract	C3	186	26	16	2	65	0	0	10	7	7	0	22	18.63	14.1
abstract.C3.11	abstract	C3	217	21	14	0	67	0	0	2	0	9	10	15	24.89	12.6
abstract.C3.12	abstract	C3	177	21	14	2	55	2	1	5	6	12	0	21	22.83	11.1
abstract.C3.13	abstract	C3	100	11	5	2	40	1	0	2	4	6	0	6	21.89	12.2
abstract.C3.14	abstract	C3	116	10	5	0	46	0	0	1	5	11	0	11	25	9.7
abstract.C3.15	abstract	C3	66	11	6	2	18	1	0	1	5	5	1	11	32.29	7.4
abstract.C3.16	abstract	C3	129	18	8	1	44	0	0	3	5	8	3	6	22.33	9.3
abstract.C3.17	abstract	C3	145	19	16	0	55	0	0	2	4	6	0	12	22.25	15

Table A.3 – Continued on next page

Table A.3 – continued from previous page

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
abstract.C3.18	abstract	C3	277	37	26	0	94	0	0	2	5	10	3	15	13.13	13.7
abstract.C3.19	abstract	C3	117	13	9	0	35	0	0	3	0	6	7	6	49	7.8
abstract.C3.2	abstract	C3	227	25	24	0	80	2	0	4	6	7	5	19	7.54	12.3
abstract.C3.20	abstract	C3	135	13	11	1	51	0	0	0	4	4	0	5	27	26
abstract.C3.21	abstract	C3	189	28	14	2	62	3	0	3	3	9	6	19	24	13.9
abstract.C3.22	abstract	C3	110	11	21	0	27	1	0	5	2	8	0	10	17.3	13.6
abstract.C3.23	abstract	C3	114	14	12	0	33	0	0	2	2	3	1	11	31	15.3
abstract.C3.24	abstract	C3	111	13	14	0	41	0	0	2	0	6	6	12	29.8	11.8
abstract.C3.25	abstract	C3	99	15	9	1	27	0	0	0	6	6	0	5	19.17	11
abstract.C3.26	abstract	C3	124	18	9	0	43	2	1	5	6	6	1	8	22.13	10.4
abstract.C3.27	abstract	C3	216	24	25	0	69	1	1	3	9	10	0	10	17.9	10.1
abstract.C3.28	abstract	C3	155	24	19	0	42	1	0	2	7	9	0	9	37	12.6
abstract.C3.29	abstract	C3	142	19	16	0	41	0	0	1	2	6	4	6	29.75	11.3
abstract.C3.3	abstract	C3	109	15	12	0	40	0	0	2	2	5	3	3	19.56	13.8
abstract.C3.4	abstract	C3	190	23	15	0	61	0	0	6	2	10	4	8	28.29	13.5
abstract.C3.5	abstract	C3	204	28	9	0	72	1	1	5	9	10	4	12	19.33	8.8
abstract.C3.6	abstract	C3	161	20	13	2	46	0	0	4	8	8	0	0	26.78	14.8
abstract.C3.7	abstract	C3	165	19	9	2	50	3	0	4	4	8	5	2	24.25	8.9
abstract.C3.8	abstract	C3	129	15	11	0	44	2	1	3	6	6	4	12	25	7.8
abstract.C3.9	abstract	C3	146	22	20	0	44	0	0	9	4	7	0	8	14.67	15.2
RA.A.1	RA	A	2369	346	155	29	609	33	3	89	165	58	1	46	19.53	9.1
RA.A.10	RA	A	4286	314	238	25	992	92	1	199	285	79	4	65	20.04	7.9
RA.A.11	RA	A	5263	670	271	27	1488	77	28	149	333	116	13	114	20.6	8.4
RA.A.12	RA	A	5637	609	417	46	1368	118	16	184	366	95	32	199	17.75	8
RA.A.13	RA	A	1258	169	100	12	330	10	0	34	85	45	6	40	23.91	8.8
RA.A.14	RA	A	15429	1869	1248	95	4075	374	69	541	1168	323	32	567	17.42	7.2
RA.A.15	RA	A	2265	232	191	31	551	44	10	83	114	60	13	112	18.77	10.2
RA.A.16	RA	A	2295	268	137	26	599	48	12	92	164	55	16	86	17.14	7.3
RA.A.17	RA	A	3002	275	341	38	660	37	14	106	161	92	6	114	22.32	11.8
RA.A.18	RA	A	5340	622	350	41	1434	132	15	173	357	104	25	172	15.27	7.6
RA.A.19	RA	A	4559	517	257	58	1229	175	16	169	286	154	31	112	15.84	9
RA.A.2	RA	A	4571	546	371	41	943	92	37	201	291	121	16	83	15.9	8.7
RA.A.20	RA	A	8867	1055	838	75	2240	184	32	323	501	177	26	334	20.02	10
RA.A.21	RA	A	4806	362	377	37	1380	83	35	159	255	104	8	229	19.22	10.6
RA.A.22	RA	A	2468	306	181	17	662	29	10	91	122	79	23	86	20.53	9.8
RA.A.23	RA	A	4664	509	303	38	1192	37	38	125	244	122	16	157	16.25	10.4
RA.A.24	RA	A	3204	350	180	7	723	52	2	151	267	63	8	37	16.74	7.1

Table A.3 – Continued on next page

Table A.3 – continued from previous page

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
RA.A.25	RA	A	3198	377	205	29	712	86	17	185	218	144	25	174	23.63	7.6
RA.A.26	RA	A	4639	598	259	51	1266	120	14	141	338	99	21	194	16.93	7.4
RA.A.27	RA	A	11644	1284	752	177	3213	185	27	458	718	300	18	317	11.57	10.7
RA.A.3	RA	A	5464	711	359	104	1310	135	14	244	347	97	23	139	21.67	8.6
RA.A.4	RA	A	5825	876	372	38	1571	101	42	184	386	185	16	89	19.67	8.1
RA.A.5	RA	A	6191	623	505	50	1488	100	17	177	376	112	24	123	16.03	8.8
RA.A.6	RA	A	3732	353	246	27	1107	79	16	91	157	119	27	181	15.56	11.8
RA.A.7	RA	A	4493	575	343	23	1160	79	15	145	279	141	23	69	16.38	8.2
RA.A.8	RA	A	6021	649	422	74	1721	82	37	202	385	222	12	231	16.12	9.6
RA.A.9	RA	A	3400	379	220	45	794	81	14	132	202	77	16	119	22.2	8.6
RA.C1.1	RA	C1	7899	1007	592	56	2050	81	9	269	459	201	40	230	23.01	9.1
RA.C1.10	RA	C1	11396	1284	1028	54	3116	100	30	367	271	333	372	659	24.93	7.8
RA.C1.11	RA	C1	8510	937	669	78	1904	106	35	360	364	236	200	473	20.06	10
RA.C1.12	RA	C1	11745	1476	1079	69	2835	213	71	520	625	248	34	449	25.09	8.2
RA.C1.13	RA	C1	1548	196	134	13	369	16	3	63	79	40	5	83	7.57	9.9
RA.C1.14	RA	C1	6350	714	477	72	1457	99	29	252	460	239	14	229	20.92	9.8
RA.C1.2	RA	C1	10901	1250	948	90	2973	171	75	422	468	263	84	386	30.14	11.5
RA.C1.3	RA	C1	12178	1436	1239	139	2866	192	44	631	642	354	133	570	23.23	9.7
RA.C1.4	RA	C1	12529	1567	984	79	3044	179	74	529	614	476	268	431	38.78	8.2
RA.C1.5	RA	C1	15467	1785	1308	172	3376	378	118	829	984	454	111	486	19.93	8.2
RA.C1.6	RA	C1	7074	888	700	62	1705	196	37	211	478	185	22	323	29.63	8.7
RA.C1.7	RA	C1	4972	517	359	42	1291	63	16	199	300	106	43	202	30.99	8.2
RA.C1.8	RA	C1	9848	1171	754	69	2537	155	71	365	356	220	262	554	25.88	8.1
RA.C1.9	RA	C1	6025	708	376	34	1534	75	8	192	299	190	111	165	25.19	8.9
RA.C2.1	RA	C2	3750	372	227	8	1099	17	9	104	83	111	80	131	21.1	13.2
RA.C2.10	RA	C2	2413	245	162	6	816	8	3	68	36	111	75	74	21.26	13.6
RA.C2.11	RA	C2	2973	375	228	7	1000	16	8	72	121	109	23	124	19.22	12.6
RA.C2.12	RA	C2	2552	228	178	7	878	11	3	39	50	94	60	72	23.46	13.8
RA.C2.13	RA	C2	4431	530	252	48	1395	37	6	135	159	170	116	141	26.22	10.4
RA.C2.14	RA	C2	4264	439	352	14	1286	17	10	105	74	161	133	173	19.88	12.2
RA.C2.15	RA	C2	4485	515	328	17	1304	43	17	156	206	243	55	184	23.08	10.4
RA.C2.16	RA	C2	3541	376	252	16	1076	24	8	113	133	158	57	78	25.21	11.9
RA.C2.17	RA	C2	2449	240	181	7	872	10	8	69	36	109	81	87	19.3	13.1
RA.C2.18	RA	C2	2665	258	210	9	944	6	0	54	72	64	48	132	23.22	13.4
RA.C2.19	RA	C2	3624	446	300	19	1117	21	3	87	98	141	99	170	21.23	11.5
RA.C2.2	RA	C2	2881	361	251	25	817	27	8	119	107	82	50	106	21.84	11.2
RA.C2.20	RA	C2	3853	436	270	11	1211	16	24	75	80	134	116	180	23.98	11.5

Table A.3 – Continued on next page

Table A.3 – continued from previous page

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
RA.C2.21	RA	C2	3119	377	256	31	966	36	9	132	156	67	30	127	25.92	11
RA.C2.22	RA	C2	2366	261	296	8	693	19	3	89	58	85	47	108	23.27	14.1
RA.C2.23	RA	C2	2750	286	167	10	958	35	2	77	45	104	100	98	22.86	11.4
RA.C2.24	RA	C2	5550	521	269	14	1667	28	16	106	50	210	210	138	22.79	12
RA.C2.3	RA	C2	4542	455	237	21	1445	32	12	121	140	196	103	191	24.68	11.6
RA.C2.4	RA	C2	1976	203	136	8	653	4	1	40	60	60	22	79	17.46	14.7
RA.C2.5	RA	C2	2687	302	187	21	832	31	9	66	77	100	44	81	23.09	13.6
RA.C2.6	RA	C2	4058	320	306	3	1399	7	9	67	64	111	74	148	22.52	17.6
RA.C2.7	RA	C2	3302	369	213	6	920	19	10	114	96	105	70	147	19.96	11.2
RA.C2.8	RA	C2	3309	378	286	33	928	33	6	137	135	81	45	118	27.55	11.2
RA.C2.9	RA	C2	2755	287	180	22	866	8	12	72	118	101	33	79	23.22	11.1
RA.C3.1	RA	C3	2936	369	202	13	809	25	6	48	82	105	61	168	21.42	11.4
RA.C3.10	RA	C3	3447	374	268	31	1031	9	2	122	175	135	6	226	21.26	12.5
RA.C3.11	RA	C3	2555	319	183	8	764	6	1	41	112	96	42	139	17.27	9.7
RA.C3.12	RA	C3	2524	256	233	31	676	16	8	79	138	116	3	180	25.6	12.9
RA.C3.13	RA	C3	2724	278	197	29	771	18	4	65	163	114	4	123	22.45	11
RA.C3.14	RA	C3	2193	253	159	15	547	19	1	61	84	72	17	135	21.37	12.3
RA.C3.15	RA	C3	2541	301	216	25	734	16	3	49	117	78	55	170	31.89	8.7
RA.C3.16	RA	C3	3608	426	285	55	1105	12	18	114	165	126	24	137	24.67	13.7
RA.C3.17	RA	C3	3304	384	271	12	1003	16	2	74	131	91	8	184	25.63	14.8
RA.C3.18	RA	C3	1965	259	178	10	609	10	2	32	53	67	17	96	22.43	16.4
RA.C3.19	RA	C3	1905	212	132	13	479	23	0	36	58	57	64	74	24.24	8.7
RA.C3.2	RA	C3	4133	483	307	17	1306	27	11	97	93	189	99	192	14.52	13.5
RA.C3.20	RA	C3	2990	354	237	30	819	19	0	92	134	119	15	102	24.12	13.3
RA.C3.21	RA	C3	2271	326	165	12	685	11	1	59	53	108	80	160	19.48	10.6
RA.C3.22	RA	C3	2212	209	267	32	488	33	6	93	123	72	3	170	26.11	13.1
RA.C3.23	RA	C3	2482	283	196	13	588	43	4	80	148	44	18	151	25.2	7.9
RA.C3.24	RA	C3	1664	196	122	11	524	7	3	32	34	44	51	84	26.74	12
RA.C3.25	RA	C3	2189	297	167	17	616	4	7	44	77	79	27	106	27.36	13.2
RA.C3.26	RA	C3	3162	423	191	11	961	23	11	85	117	114	52	169	26.34	10.7
RA.C3.27	RA	C3	2973	328	251	16	822	24	7	80	137	97	15	151	21.64	11.4
RA.C3.28	RA	C3	3797	523	272	27	1161	12	1	77	176	154	17	209	22	12.1
RA.C3.29	RA	C3	3770	499	257	30	1119	23	4	64	93	177	153	205	22.44	9.7
RA.C3.3	RA	C3	2475	305	167	9	814	5	6	55	46	96	74	87	27.11	12.4
RA.C3.4	RA	C3	2313	264	221	7	690	10	5	52	108	94	16	95	19.93	11.4
RA.C3.5	RA	C3	2283	260	159	11	783	10	4	63	97	78	16	81	25.09	12.1
RA.C3.6	RA	C3	3517	395	271	38	1008	12	7	104	144	133	34	104	23.41	13.2

Table A.3 – Continued on next page

Table A.3 – continued from previous page

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
RA.C3.7	RA	C3	1797	207	108	17	504	15	3	51	63	87	53	43	29.37	9.6
RA.C3.8	RA	C3	3491	368	222	29	1061	25	3	108	168	121	17	234	19.38	12.4
RA.C3.9	RA	C3	2177	258	190	25	642	7	1	64	106	86	18	101	23.31	12

Table A.3: Data for inductive analysis

A.6 Script for inductive analysis in R

This section describes the script used in R for the hierarchical agglomerative cluster analysis and for the principal component analysis. Each of the commands is explained in a comment-line initiated by the symbol #.

```
Daten <- read.csv2("mydata.csv")
```

# data matrix (cf. Table A.3) is loaded in R

```
Daten.1 <- Daten[-61,]
```

# text *abstract.C2.5* is deleted from the matrix for it being misbuilt

```
M <- Daten.1[,-(1:4)]
```

# only the numeric columns are taken in the matrix. New matrix is called M.

```
lexical.density <- Daten.1$lexical.density/100
```

# vector lexical.density is created containing the values of lexical.densitiy divided per 100, so that the final values in the scaled and centered matrix are all between 0 and 1

```
sentence.length <- Daten.1$sentence.length/100
```

# vector sentence.length is created containing the values of sentence.length divided per 100, so that the final values in the scaled and centered matrix are all between 0 and 1

```
library(MASS,rpart,amap)
```

# required libraries are loaded

```
text.sizes <- Daten.1$tokens
```

# vector containing the total amount of texts per text is created. It is used later for the normalization of the matrix.

```
rownames(M) <- Daten.1$text
```

# each row receives its names from the name of the corresponding anonymized texts.

```
M.r <- M / text.sizes
```

# matrix data is normalized. New matrix is called M.r

```
M.r$sentence.length <- sentence.length
```

# the column sentence.length is replaced by the former defined sen-

```
tence.length
M.r$lexical.density <- lexical.density
# the same for lexical.density
# Now M.r is a scaled matrix
write.csv2(as.matrix(M.r[,]), file="M-r.csv")
# M.r is exported in a csv file
M.s <- scale(M.r)
# M.s is the scaled matrix M.r. Scale is a function in R whose default method centers and/or scales the columns of a numeric matrix.
write.csv2(as.matrix(M.s[,]), file="M-s.csv")
# the scaled matrix M.s is exported in a csv file)
distances <- dist(M.s)
# the distance matrix is generated using the function dist in R, which computes the distances between the rows of a matrix. This is the matrix used in cluster, PCA and mds.
clusters <- hclust(distances, method="complete")
# hierarchical agglomerative cluster analysis is performed
plot(clusters)
# the cluster is plotted
M.s.t <- t(M.s)
# M.s is inverted for clustering for the features
distances.t <- dist(M.s.t)
# the corresponding distance matrix is generated
clusters.t <- hclust(distances.t, method="complete")
# hierarchical agglomerative cluster analysis is performed
plot(clusters.t)
# the cluster for the features is plotted
rect.hclust(clusters, k=2, border="red")
# two rectangles are plotted separating the main clusters
M.r <- as.matrix(M.r)
```

```
#rpart(), function for classification trees requires data in a matrix, not data.frame

tree.type <- rpart(Daten.1$type ~ M.s, method="class")

plot(tree.type)

text(tree.type, cex=0.6, use.n=TRUE, all=TRUE)

# these last three command lines generate and plot the classification tree for text types, i.e., abstracts and RAs

tree.domain <- rpart(Daten.1$domain ~ M.s, method="class")

plot(tree.domain)

text(tree.domain, cex=0.6, use.n=TRUE, all=TRUE)

# these last three command lines generate and plot the classification tree for domains, i.e., computer science, linguistics, biology, and mechanical engineering

pairs(M.s, col=as.integer(Daten.1$type), pch=3)

pairs(M.s, col=as.integer(Daten.1$domain), pch=3)

# these last two command lines generate the pair plots for text types and domains

M.pca <- prcomp(M.s)

# principal component analysis (PCA) is performed

plot(M.pca)

biplot(M.pca)

# PCA is plotted

mds <- isoMDS(distances)$points

plot(mds, cex=1.6, col=as.integer(Daten.1$domain), pch=20)

plot(mds, cex=1.2, col=as.integer(Daten.1$type), pch=20, lwd=1.4)

# these last three command lines generate the multidimensional scaling and plot the corresponding graph according to domain and to text type, respectively

library(scatterplot3d)

cl <- ifelse(Daten.1$type=='abstract','black','red')
```

# colors are defined, if abstract then black, otherwise red

```
scatterplot3d(M.pca$x[,1], M.pca$x[,3], M.pca$x[,2],
color=cl, pch=20, xlab="PC1", ylab="PC3", zlab="PC2")
```

# the 3d scatterplot is produced

```
legend(locator(1),c("Abstracts","RAs"),pch=c(20,20),col=c(1,2),
cex=0.8, bg="white")
```

```
cl2 <- as.integer(Daten.1$domain)
```

# colors are defined: computer science = 1, linguistics = 2, biology = 3; mechanical engineering = 4

```
scatterplot3d(M.pca$x[,1], M.pca$x[,3], M.pca$x[,2],
color=cl2, pch=20, xlab="PC1", ylab="PC3", zlab="PC2")
```

```
legend(locator(1),c("Computer science","Linguistics",
"Biology",
"Mechanical Engineering"),pch=c(20,20,20,20,20),col=c(1,2,3,4),
cex=0.8, bg="white")
```

Erklärung

Die vorliegende Arbeit wurde von mir selbständig verfasst. Die zur Bearbeitung des Themas herangezogenen Quellen, die Literatur und sonstige Hilfsmittel wurden entsprechend gekennzeichnet.

Es wurde von mir noch kein Promotionsversuch, auch nicht an einer anderen Universität, unternommen.

Darmstadt, den 10 Januar 2011.

Curriculum Vitae

Mônica Holtz

PhD, English studies
July 2006 – December 2010
Technische Universität Darmstadt
Darmstadt, Germany

Magistra Artium (M.A.), German & English studies, Chemistry
April 2000 – June 2006
Technische Universität Darmstadt
Darmstadt, Germany

Master of Science (MSc.), Organic Chemistry
April 1990 – November 1992
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil

Bachelor (BA), Chemistry
March 1985 – March 1990
Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro, Brazil