UNIVERSITY OF JYVÄSKYLÄ

TEXT TYPE DISTINCTIONS AND VARIATION IN ENGLISH OF SOFTWARE ENGINEERING

A Pro Gradu Thesis in English

by

Riitta-Liisa Hämäläinen

Department of Languages

2008

HUMANISTINEN TIEDEKUNTA
KIELTEN LAITOS

Riitta-Liisa Hämäläinen
TEXT TYPE DISTINCTIONS AND VARIATION
IN ENGLISH OF SOFTWARE ENGINEERING

Pro gradu -tutkielma

Englannin kieli
Toukokuu 2008 79 sivua

Tieteen ja tekniikan englannin kielen tutkimuksessa on perinteisesti keskitytty kielen syntaktisiin rakenteisiin ja retorisiin keinoihin. Tämän tutkielman tarkoituksena on selvittää, onko eri vastaanottajaryhmille kirjoitettujen ohjelmistotekniikan (software engineering) tekstien englannin kielessä variaatiota. Lähtökohtana on voidaanko tieteellistä englannin kieltä tutkia erilaisten kielellisten piirteiden pohjalta, ja voidaanko niitä käyttää myös laajemmin eri tekstityyppien vertaamiseen ja tieteellisen englannin kielen yleisen kuvauksen kehittämiseen.

Viitekehyksenä on Widdowsonin (1979) näkemys tieteellisestä kielestä, jonka mukaan yleistieteellisten lehtien kieli eroaa tieteen opetuksen ja tutkimuksen kielestä ja lähestyy piirteiltään ei-tieteellistä kieltä. Tutkimusasetelmassa ohjelmistotekniikan kieli jaetaan kolmeen tekstityyppiin: tieteelliset tutkimusartikkelit, yliopiston ensi vaiheen oppikirjat ja tavallisille ihmisille kohdistetut tietotekniikan lehtiartikkelit. Tekstien kohderyhmät ovat vastaavat. Tutkimusmateriaalina on 8 tekstikatkelmaa jokaisesta tekstityypistä.

Tutkimuksessa vastataan kysymyksiin: 1) Voidaanko hypoteettiset tekstityypit erottaa toisistaan? 2) Ilmeneekö eri kohderyhmille kirjoitettujen ohjelmistotekniikan tekstien välille variaatiota?

Tutkielmassa käytetään Biberin (1988) monidimensionaalista kielellisen variaation tutkimusmenetelmää. Sen mukaan usein yhdessä esiintyvillä kielellisillä piirteillä on yhteisiä kommunikatiivisia funktioita, joista voidaan muodostaa kielellisiä dimensioita. Näiden dimensioiden perusteella voidaan vertailla eri tekstityyppejä ja niiden variaatiota. Tutkimusmenetelmässä variaatioanalyysiin yhdistyy empiirinen kvantitatiivinen tutkimusote.

Tekstityypit voitiin eriyttää toisistaan. Lehtiartikkelien tekstityyppi eriytyi jossain määrin kahden faktorin ja oppikirjojen tekstityyppi yhden faktorin perusteella muista tekstityypeistä. Kielellisten piirteiden kommunikatiivisten ominaisuuksien perusteella määriteltyjen dimensioiden pohjalta variaatiota voitiin osoittaa jonkin verran lehtiartikkelien tekstityypin ja kahden muun tekstityypin välillä yhden dimension suhteen. Muiden dimensioiden yhteys variaatioon jäi puutteelliseksi.

Asiasanat: text analysis. ESP. scientific discourse. linguistic analysis. text type. variation.

TABLE OF CONTENTS

List of figures and tables

1 INTRODUCTION

Discourse in various special fields is different. You notice it, for instance, when you need to describe some scientific phenomenon or technical equipment in English and would like to do it as a native speaker of English would. And not as any native, but one who has knowledge of the special field in question, so that you can be sure that the expressions are understandable and idiomatic scientific or technical English, which English speaking experts perfectly understand.

But how is the English of technology and science different from everyday English? The issue has been studied widely in the field of English for special or specific purposes, ESP, or English for science and technology, EST. In earlier studies the focus has ranged from grammatical and syntactical issues to questions of rhetorical choices and politeness strategies. Even though they have given valuable information of scientific English, the characteristics of ESP or EST still need further study, to be able to define text comparability and reach an overall description of scientific English.

This study examines text type variation of scientific texts based on their linguistic features. It combines textual and variation analysis with quantitative measures, which has not been a common approach in ESP. The language of software engineering, the topic of this study, seems not to have been studied earlier. The purpose of this study is to find out if there is variation in the scientific text types of software engineering written for different audiences.

The theoretical framework of this study is adopted from Widdowson’s (1979) views on the description of scientific discourse; the discourse of scientific journalism has different rhetorical principles from those of scientific instruction and exposition and it approaches non-scientific discourse.

For the purposes of this study the scientific language of software engineering is divided into three groups of text types; academic research articles, introductory textbooks, and popular journals. The audience is divided likewise into three groups; academic, introductory university, and laypersons.

The method of this study, the multi-dimensional (MD) analysis of linguistic variation and the concept of dimension of linguistic variation, is adopted from Biber (1988). It suggests that linguistic features which occur frequently together in texts, for instance, verbs, nouns, and pronouns, have at least one shared, underlying communicative function, and that the functions can be linked to variation in the communicative situation. The frequencies of the co-occurring features in the texts are grouped into linguistic factors by statistical methods. The texts and their relations are compared with respect to these factors, which are then interpreted as linguistic dimensions, by assessing the communicative functions of the features that constitute them. The dimensions are given names, labels, to indicate why the features co-occur in various text types. The co-occurring linguistic features and their linguistic dimensions offer a way to discuss variation in text types.

The study is structured so that Chapter 2 begins with a short history of ESP/EST with the theoretical framework of this study adopted from Widdowson’s (1979) view on scientific discourse, followed by a review of previous studies on analysing ESP/EST. The chapter continues then with a description of the method of this study, the MD analysis, with a review of other studies where the same method of analysis has been used.

Chapter 3 begins with the research design of this study with research questions, and it is followed, firstly, by the introduction of the data, the sample texts and the linguistic variables and their coding, and, secondly, by an overview of the statistical analysis and extraction of four linguistic factors, the final factorial structure. The chapter continues to establish the findings of the analysis, based on the frequencies of the linguistic features and the factorial structure. Firstly, I analyse the frequencies of the linguistic features in the corpus of this study and their statistics in some detail, as they are used in the analysis to answer the two research questions. Secondly, I survey the relations of the text types with one another regarding the four extracted factors, to answer the first research question, *Can the hypothesised text types be distinguished?* Thirdly, to answer the second research question, *Do software engineering texts written for different audiences show variation?,* I assess the communicative functions shared by the features grouped on each factor, with reconsideration of the relations of the text types, and suggest labels for the linguistic dimensions to explain why the features co-occur and show variation.

In Chapter 4 I summarise the purpose and method of the study with the main results, compare the results with those of other studies, evaluate the study, and suggest further research.

2 BACKGROUND

Without going into further details concerning the terminology within ESP, for the purposes of this study ESP is used throughout the study to refer to the use of English in special fields such as, for instance, software engineering, which links it closely to English for science and technology.

As a research context to this study, approaches to describing ESP/EST will be reviewed with historical perspectives in section 2.1. The theoretical framework of this study, Widdowson's view on scientific discourse, will be introduced in section 2.2. Previous studies on analysing language in ESP will be surveyed in section 2.3, and the method of this study, multi-dimensional linguistic analysis, presented in section 2.4. Section 2.5 reports studies which have applied the MD method of linguistics analysis.

2.1 Approaches on describing ESP with historical perspectives

The concept of *English for special or specific purposes*, ESP, is far from being straightforward. Robinson (1991: 1), for instance, suggests that ESP involves education, training and practice, drawing upon many realms of knowledge such as language, pedagogy and students'/participants' specialist areas of interest, and claims that it is impossible to give a universally applicable definition of ESP. ESP is divided into many types and acronyms, of which a major distinction is drawn between *English for occupational purposes,* EOP, which involves work-related needs and training, and *English for academic purposes,* EAP, which is related to academic study needs (Robinson 1991: 2). *English for science and technology*, EST, cuts across these, being mainly used for ESP in the USA, and to refer both to work- and study-related needs (Robinson 1991: 2).

A short review of ESP reveals that there are several approaches to its history as well as to its starting points.

According to Strevens (1977: 146), ESP is a case of special purpose language teaching, and he proposes that courses for travellers have been one of the first instances of it, suggesting further that ESP started in the first half of the 20th century as a pedagogical means to fulfil the specialised needs of various English learning groups, such as tourists, students and employees of special fields. Teaching materials such as bilingual dictionaries and phrase books intended especially for travellers date back to around the 1930s and foreign language courses for science students have an even longer history; in these courses grammar was taught with texts taken from the required field of science and the courses aimed at teaching the student to translate some basic texts of their special field with the help of a dictionary (Strevens 1977: 150-151).

The Second World War was an important stage in the history of ESP/LSP, because groups of people had to be taught a sufficient command of a foreign language in order to cope with their tasks connected with war operations (Strevens 1977: 151). Hutchinson and Waters (1987: 6-7) also regard the war as a starting point of ESP. Firstly, after the war scientific, technical and economic activities became international and English became the international language of technology and commerce. Secondly, the ability to use English in these special fields was more and more required especially after the war, when the English language became increasingly used with the widening global trade, resulting in increased numbers of people wishing to have English courses developed to meet the specific needs of their working environment. It was thought that a group of learners of English should be taught the linguistic features of their special area, whether it be technical, commercial, or some other field of work or study (Hutchinson and Waters 1987: 8).

Another factor having an effect on the rise of ESP was that in educational psychology the attitude of the learner, especially his or her motivation to learn, was emphasised as a key to successful learning (Hutchinson and Waters 1987: 8). So, in order to motivate learners, materials in English courses for special areas were taken from their special fields. Johns and Dudley-Evans (1991: 298-299) are of the same opinion, stating that ESP examines written or oral language in order to provide suitable and as good as possible pedagogical materials for the needs of language learners within a special learning context.

Quite a different view on the history of ESP is that of Halliday’s (1988), who studied the development of scientific English from a historical perspective, referring by scientific English mostly to technical nouns and nominalizations. He studied Newton’s *Optics, or a Treatise of the Reflections, Refractions, Inflections and Colours of Light* (published in 1704; written in 1675-87), which he regards as the starting point of scientific English. Halliday (1988: 171-172) claims that scientific English developed along with the development of experimental scientific methods, the reporting of which required new kinds of rhetorical means and logical argumentation. Nominal elements increased and took over semantic content in clauses. Atkinson (1992: 337) considers that the beginning of scientific journal was of decisive importance for the development of scientific English.

Swales (1985: xiv) suggests that EST/ESP started in 1962 when an article “Some Measurable Characteristics of Modern Scientific Prose” by C. L. Barber was published. Barber’s study was a frequency analysis of sentence length, clause types, verb tenses, and the uses of non-finite verbs, as well as of vocabulary, attempting to identify the characteristics of scientific English (Barber 1962, as quoted by Swales 1985: 4).

2.2 Widdowson’s view on scientific discourse

Widdowson (1979: 19-20) characterises EST from the following scientific discourse and pedagogical perspectives. Firstly, EST is a textualisation of a variety of discourse which is independent of any language and which expresses a secondary and universal culture which those involved with science and technology learn during their education. Secondly, when EST is regarded as varieties of discourse textualised in English, in contrast to other textualisations in other languages, in a teaching situation the learner can use the knowledge of the discourse already learned in his/her area of science and technology in his/her first language and use that knowledge for learning the particular textualisation of this discourse in English. The notions of scientific discourse being universal communication, processed by textualisation and realised by scientific text in different languages, are shown below in Figure 1:

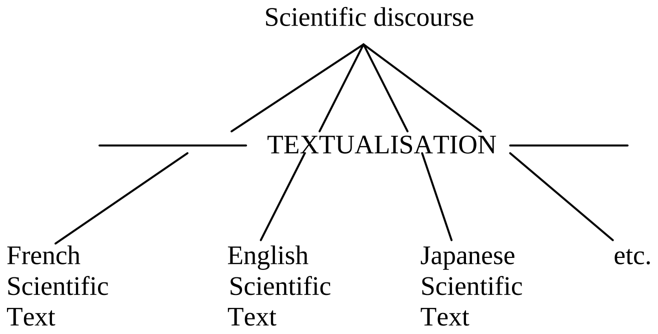


Figure 1. Scientific discourse, universal for all languages (adapted from Widdowson 1979: 52)

Widdowson (1979: 52-53) further describes scientific discourse as follows. Firstly, the discourse of scientific instruction, as in text books, is different from scientific exposition, as in research papers. During the different levels of scientific instruction students learn the scientific concepts and procedures of their discipline and at the same time the rhetorical principles of scientific discourse. In early scientific instruction the earlier experiences of learners play a part and the instruction is related to their primary cultures, but during later stages of instruction the secondary culture of science develops and little by little learners become scientists. Secondly, scientific journalism is rhetorically very different, because it relates the findings of the secondary culture of science into primary culture terms. Thus the discourse of scientific journalism has different rhetorical principles from those of scientific instruction and exposition and it approaches non-scientific discourse.

According to Widdowson’s rhetorical scale (1979: 53), scientific discourse includes different kinds of communication as presented in Figure 2 below:

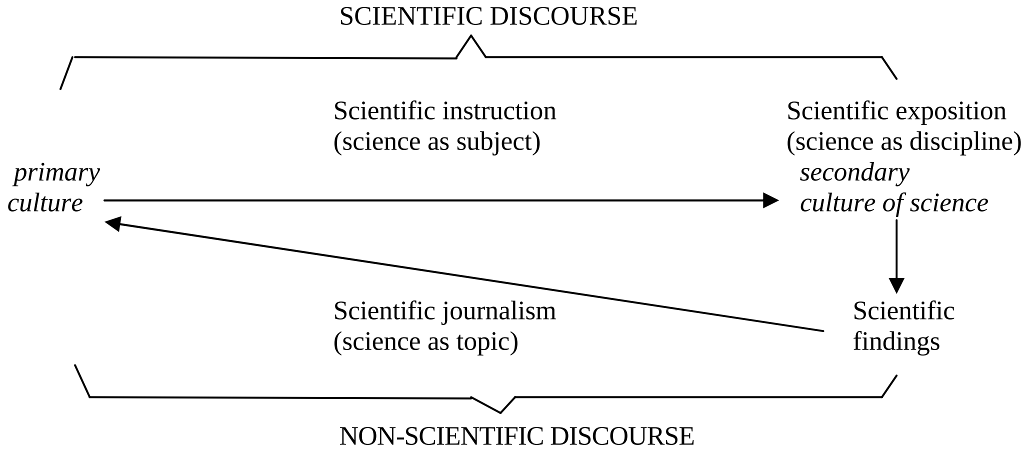


Figure 2. Scientific vs. non-scientific discourse (adapted from Widdowson 1979: 53)

2.3 Previous studies on analysing language in ESP

Barber's study in 1962 mentioned above, as quoted by Swales (1985: xiv), is said to have been one of the first attempts to use statistical methods in identifying the characteristics of scientific English, which was new at that time. Barber's corpus included three texts, the first one from a university textbook on electronics engineering applications, the second one concerning basic research in the field of biochemistry, and the third one from an elementary university textbook on astronomy. Barber studied sentence length, clause types, verb tenses and the uses of non-finite verbs, and his results were mainly presented in numbers and percentages of their occurrences. In Barber's material the number of progressive forms was found to be very small.

Lackstrom, Selinker and Trimble (1973) studied the effect of rhetorical choices on grammatical forms, such as article use and tense choice, which they had found difficult to describe linguistically and to teach to foreign students. For them the paragraph is the basic unit of discourse in EST, and their concept of conceptual paragraph, which includes sentences that form a complete generalisation, thought, may cover one or more physical paragraphs. Accordingly, the writer, by making rhetorical decisions, organises information and relates concepts for particular readers to make it easier for them to understand the concepts and meaning of the text. Lackstrom, Selinker and Trimble suggest that the choices the writer makes in developing the patterns of organizational development have an effect on their grammatical choices, for instance, the choice of tense, in relation to the time framework of reporting. They also claim that the writer selects the article to go with a noun phrase on the basis of the degree of generality. Lackstrom, Selinker and Trimble have a wider perspective on language analysis with their concept of conceptual paragraph, than, for instance, Barber mentioned above, but their focusing on only some grammatical choices, such as use of articles and tense choice, was, however, rather limited.

Other linguists that have taken account of the rhetorical point of view, the organizational and discoursal structure of a text, have been Tarone et al. (1981) in their statistical study, which examined use of the passive and active voice in two journal articles on astrophysics. The scholars Tarone et al. wanted to check if the general opinion concerning the passive voice being very typical in EST was true or not, and they studied the occurrence of passive and active verb forms and the rhetorical function of the passive. The number of passive and active verb forms was counted and an informant of astrophysics, one of the writers of the study, outlined the rhetorical structure of the articles. Tarone et al. came to the conclusion that the active voice, and especially the active first person plural *we*, was used much more frequently in these articles than the passive.

Swales (1981) conducted a study of bare attributive –*en* participles, which have no clause, phrase, or modifying adverb attached to them at the surface structure level. His aim was to study the factors that make the non-native speaker of English to use or not to use a bare participle and where to place it. Swales came to the conclusion that the use of pre- and post-posed bare participles is largely determined by semantic reasons; for instance, *the stolen jewels* and *the jewels stolen* or *added heat* or *heat added* can convey different meanings.

A study of quite another type within ESP was that of Strother and Ulijin (1987: 99-100), who were interested in the question of whether to use authentic or rewritten texts in ESP pedagogy. They studied if syntactic rewriting affects comprehension of EST texts. In this study native speakers and non-native speakers of English, students of computer technology and humanistic sciences, read authentic and rewritten texts and their text comprehension was tested. Strother and Ulijin's conclusion was that syntactic simplification does not increase text comprehension.

Beedham and Bloor (1989) studied the relationship between communicative function and lexico-grammatical form in English for computer science. Their material consisted of three textbooks of computer education, in which they identified thirteen communicative functions, such as, for instance, description of algorithms, commentary on programs, comparison, history, definitions, analogies, exemplifications, introduction of problem, and calculation. They concentrated in the communicative function of the description of algorithms and its formal realisations. An example of description of algorithms: *Search to the right until we find the first stack j (>i) which is not using all its storage space, i.e. Top [j]<Bottom [j + 1]*. The results of their study were two-fold. Firstly, Beedham and Bloor (1989: 16) claimed that the formal realisation of the communicative function is not grammatical, because grammatical forms such as, for instance, imperative, conjunctions, passive, infinitive, present tense, modals, progressive, and present perfect, appeared throughout the thirteen communicative functions. Secondly, Beedham and Bloor (1989: 20) suggested that lexical items such as *algorithm*, *describe, method*, *operation*, *traversal*, as a name of algorithm, *searching*, as a name of operation, *technique*, were explicit realisations of communicative functions. In addition, they claimed that there are also numerous other lexical items, mainly of technical vocabulary that function as implicit lexical markers of the communicative function.

Politeness strategies have been an object of wide interest within ESP. Myers (1989) studied pragmatics of politeness, claims and denial of claims, in written research articles of molecular biology. Myers (1989: 30) suggests that even though scientific writing does not involve direct personal contact between the writer and the audience, it is interaction that makes the scientific writer use complicated forms for suggesting shared assumptions even in criticism and personal attributes to hedge claims or to criticise.

In the studies reported above on analysing scientific English the focus has ranged from grammatical and syntactical issues to questions of rhetorical choices and politeness strategies. Even though they have given valuable information of scientific English from their specific angle, the specific features of ESP or EST still need further study. This study examines text type variation of scientific texts, to characterise and describe scientific text types based on their linguistic features. Such characterisations and descriptions will be useful for defining text comparability and reaching an overall description of scientific English.

The study combines textual and variation analysis with quantitative measures, which has not been a common method in ESP, and the topic is the language of software engineering, which seems not to have been studied earlier. The purpose of this study is to find out if there is variation in the scientific text types of software engineering written for different audiences.

For this kind of analysis the method is provided by the multivariate statistical and functional analysis, which is also called the multi-feature/multi-dimensional, MF/MD, approach. In this study initials MD are used to refer to the multi-dimensional method of linguistic analysis described in more detail in the next section.

2.4 Multi-dimensional linguistic analysis (MD)

The methodology and main concepts of the MD method of linguistic analysis of text type (also called linguistic or register) variation will be shortly described here for the purposes of this study. The method is based on Biber’s (1988: 121) study of the relations between spoken and written English, where he attempted to specify linguistic similarities and differences among English texts (referring by text also to extracts of spoken English).

The main characteristics of the MD approach are as follows (Biber 1988: 24). Firstly, the MD approach to linguistic variation is understood to be multi-dimensional, no single dimension being enough to cover complex linguistic variation. Secondly, the dimensions are continuous scales of variation, instead of being dichotomous poles. Thirdly, the co-occurrence relations among the linguistic features of the texts are discovered through quantitative methods and statistical techniques.

Biber (1988: 63-64) suggests that the linguistic features that occur frequently together in texts, for instance, verbs, nouns, and pronouns, have at least one shared, underlying communicative function and that the functions of these linguistic features can be linked to variation in the communicative situation. In Biber’s method the frequently co-occurring linguistic features are first identified in the corpus by quantitative methods. The linguistic features are pre-determined on the basis of earlier research, and the numbers of their occurrence in the texts are calculated and the frequencies are grouped into linguistic factors by statistical methods.

For his study Biber (1988: 72) selected, based on earlier research, 67 linguistic features which had been described as functional markers in texts and which came from sixteen grammatical categories such as, for instance, tense and aspect markers, place and time adverbials, pronouns and pro-verbs, questions, nominal forms, passives, stative forms, subordination features, prepositional phrases, adjectives and adverbs, lexical specificity, lexical classes, modals, and specialised verb classes. A detailed list of the linguistic features can be found in Biber’s (1988: 73-75) Table 4.4.

The computerised text corpora in Biber’s (1988: 66-67) study comprised texts of a total of 23 genres, such as, for instance, press reportage, editorials, press reviews, biographies and essays, official documents, telephone conversations, radio broadcasts, speeches, and letters. The approximate total number of words of the corpora added up to 960,000. Biber (1988: 72) used a computational tool, a tagging program, for marking the linguistic features, and multivariate statistical techniques such as factor analysis, to count the frequencies of the linguistic features to identify sets of features that co-occur in texts. In the factor analysis a large number of original variables, frequencies of linguistic features, was clustered to a small set of derived variables, factors (Biber 1988: 79).

The factors were interpreted by assessing the communicative functions of the features constituting each factor and given names, labels, to indicate why the features co-occur in various text types and to identify linguistic dimensions underlying the factors (Biber 1988: 101). The texts types and their relations were also compared with respect to the dimensions specified by computed factor scores (Biber 1988: 63-64). Before the interpretation the factors were rotated, to enable each linguistic feature load only on one factor, and also to enable each factor to be characterized by the relatively few features that are most representative of the underlying construct (Biber 1988: 104).

Table 1 below is a summary of Biber’s (1988: 89-90, 120) six factor structure with the interpretative labels given for the factors, or linguistic dimensions, of linguistic variation.

Table 1. Summary of the factorial structure in Biber (1988: 89-90, 122)

FACTOR 1
'Involved versus Informational Production'
private verbs
THAT deletion
contractions
present tense verbs
2nd person pronouns
DO as pro-verb
analytic negation
demonstrative pronouns
general emphatics
1st person pronouns
pronoun IT
BE as main verb
causative subordination
discourse particles
indefinite pronouns
general hedges
amplifiers
sentence relatives
WH questions
nonphrasal coordination
WH clauses
final prepositions
(adverbs)
(conditional subordination)
- - - - - - - - - - - - - - - - - -
nouns
word length
prepositions
type/token ratio
attributive adjs.
(place adverbials)
(agentless passives)
(past participial WHIZ deletions)
(present participial WHIZ deletions)

FACTOR 2
'Narrative versus Non-narrative Concerns'
past tense verbs
third person pronouns
perfect aspect verbs
public verbs
synthetic negation
present participial clauses
- - - - - - - - - - - - - - - - - -
(present tense verbs)
(attributive adjs.)
(past participial WHIZ deletions)
(word length)

FACTOR 3 'Explicit versus Situation-Dependent Reference'
WH relative clauses on object positions
pied piping constructions
WH relative clauses on subject positions
phrasal coordination
nominalisations
- - - - - - - - - - - - - - - - - - - -
time adverbials
place adverbials
adverbs

FACTOR 4
'Overt Expression of Persuasion'
infinitives
prediction modals
suasive verbs
conditional subordination
necessity modals
split auxiliaries
(possibility modals)
- - - - - - - - - - - - - - - - -
- - no negative features - -

FACTOR 5
'Abstract versus Non-Abstract Information'
conjuncts
agentless passives
past participial clauses
BY-passives
past participial WHIZ deletions
other adverbial subordinators
(predicative adjs.)
- -
(type/token ratio)

FACTOR 6
'On-Line Informational Elaboration'
THAT clauses as verb complements
demonstratives
THAT relative clauses as adj. complements
(final prepositions)
(existential THERE)
(demonstrative pronouns)
(WH relative clauses on object positions)
- -
(phrasal coordination)

Even though some linguistic features have loadings on more than one factor, to ensure the independence of the factor scores, each feature is included only once in the factor score of the factor on which it has the highest loading, whether it is a minus or plus value (Biber 1988: 93). When a feature is marked in parenthesis, it means that its loading is not used in the calculation of the factor scores of that particular factor, but in some other factor where it has a higher loading. The linguistic features with positive loading scores are on the top above the dotted line, and the features with negative scores on the bottom below the dotted line. The positive and negative loadings of the features on a factor occur in texts in a complementary pattern (Biber 1988: 88). In other words, when a text includes a high frequency of features with a positive loading, the features with a negative loading occur more infrequently or are non-existent in that text, and vice versa.

The dimensions of linguistic variation refer, according to Biber (1988: 9), to situational or functional parameters, such as formal / informal, literary / colloquial, and restricted / elaborated. He calls them dimensions, instead of parameters, because they define continuums, instead of being discrete poles. A text can be more or less formal and the formal/informal axis can be considered a continuous dimension of variation.

As an example of Biber's (1988: 89) interpretation of factors as textual dimensions we can have a look at Factor 2 and see how it was interpreted. The linguistic features of Factor 2 are repeated here from Table 1:

Factor 2
past tense verbs
third person pronouns
perfect aspect verbs
public verbs
synthetic negation
present participial clauses

(present tense verbs)
(attributive adjectives)
(past participial WHIZ deletions)
(word length)

Biber (1988: 108-109) suggests, for instance, that in this factor the linguistic features with positive weights, that is, the ones above the dotted line, are markers of narrative discourse. He suggests that narrative discourse includes past events, described by past tense and perfect aspect verbs, as well as reference by third person personal pronouns to typically human referents apart from the speaker and addressee. Public verbs, such as *admit, assert, declare, hint, report, say*, function as markers of reported speech. The complementary distribution of present and past tense verbs is well understandable; either past or present events are typically reported in discourse, but they are not mixed. The co-occurrence of attributive adjectives and present tense verbs signify a more frequent use of elaborated nominal referents in discourse of non-narrative types than in narrative ones. Thus this linguistic dimension is labelled by Biber 'Narrative versus Non-narrative Concerns'.

The similarities and differences among the genres of the corpus were then considered regarding the mean scores of each of the dimensions in Table 1 (Biber 1988: 121). Figure 3 below shows an example of the relationship of all genres with Biber's Dimension 2, 'Narrative versus Non-narrative Concerns'.

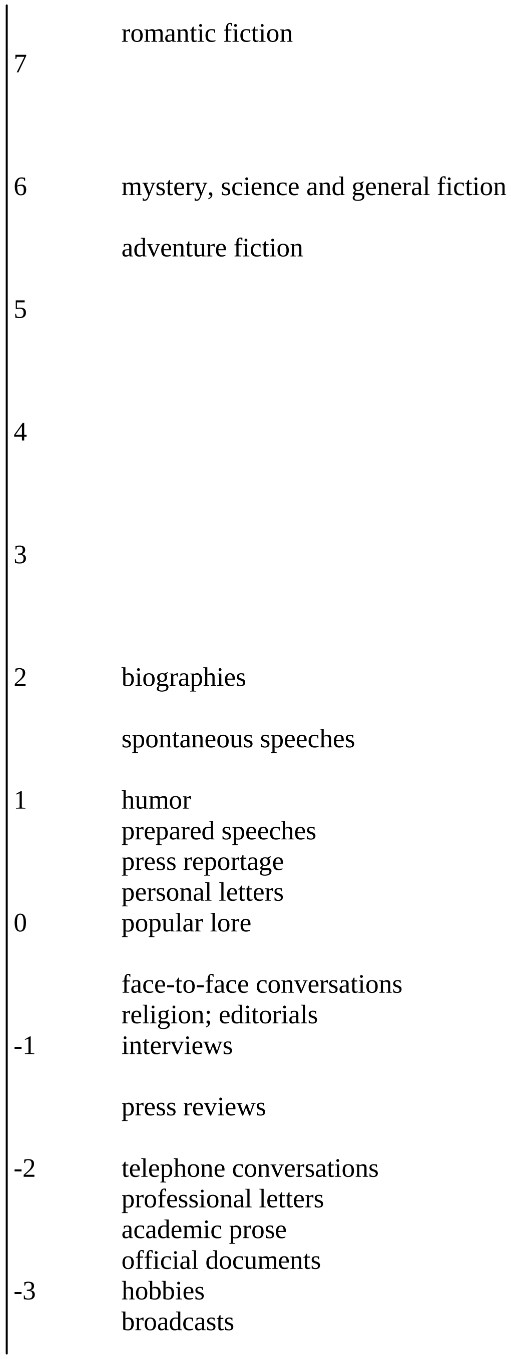


Figure 3. Mean scores of Dimension 2 for each of the genres (in Biber 1988: 136)

On this dimension fiction genres have high mean scores, and they are characterized, for instance, by frequent occurrences of past tense and perfect aspect verbs, third person pronouns and public verbs, having fewer occurrences of present tense verbs and attributive adjectives, whereas genres with low scores on this dimension have the opposite characteristics (Biber 1988: 135-

136). Biber (1988: 137) suggests that the interpretation of a narrative versus non-narrative dimension is an accurate description of the underlying function.

The result of Biber’s study concerning the differences between speech and writing was not simple and straightforward, but of a more complex nature. Biber suggests (1988: 160) that no absolute difference between speech and writing could be found, because written and spoken texts overlap with respect to each dimension. He (1988: 168) further claims that the relationship between any two genres has to be based on consideration of all six dimensions and so the characterisations of similarity or difference with respect to single dimensions are not enough.

2.5 Other studies applying the MD method

A number of studies have been conducted by using the MD approach to multivariate statistical and functional analyses of text type variation in English. As the MD analysis is based on computational and statistical tools, it provides additional empirical information to earlier studies.

Grabe (1987) conducted a study on expository prose firstly to find out if it could be defined as a text genre, secondly if it had sub-types and what their characteristics would be, thirdly to find a way to define text comparability in two or more languages, and fourthly to find means for research in contrastive rhetoric. The corpus contained texts from academic and general journals, textbooks, newspaper editorials and local news stories, which have in earlier studies been referred to, according to Grabe (1987: 117), as expository prose, and some fictional narratives and professional correspondence texts were used as control group texts. In setting up the text type groups, Grabe (1987: 117) hypothesised two sociolinguistic parameters, topical context and audience. The topical context was expected to vary along a parameter of natural science, social science and humanities, and audience along a parameter of academic, introductory university and popular. The corpus consisted of 150 texts, a total of about 70,000 words. In the study 27 syntactic variables, such as prepositions, pronouns, nominalizations, tenses, passive, relative clauses, contractions, etc., were counted. In addition, six cohesion variables, such as definite article reference, deictic reference, repetition, lexical inclusion, comparatives and synonymy/antonymy, were included in the study and counted by hand.

The measurement procedures and statistical analysis were carried out employing the MD approach (Grabe 1987: 119-120). As a result of his study, Grabe (1987: 126-128) found support for his labelling of four textual dimensions which discriminate text types: 1) ´Immediacy of context´ (immediate vs. distant context), 2) ´Interactive – informational orientation to discourse´, 3) ´Logical vs. situational context´ (information type abstract/logical vs. situational), and 4) ´Presentation of information´ (objective/effaced vs. expressive). Grabe suggests (1987: 127-128) that Factors 3 and 4 also verify his hypothesised sociolinguistic parameters. Firstly, Grabe claims (1987: 127-128) that Factor 3, information type, showed a clear separation by factor scale scores for the text types along the academic – popular parameter, and secondly, that Factor 4, ´Presentation of information´, verifies that contextual differences of natural science vs. social science vs. humanities show distinctions which separate different text types. Grabe’s results (1987: 128) suggest that expository prose is a genre with identifiable sub-types. Grabe (1987: 132-133) also conducted a cluster analysis, which allowed him further to suggest that general expository prose is a distinct text type/genre, as opposed to non-expository texts such as narratives and correspondence.

Another study to make use of the MD approach was that of Atkinson (1992) concerning the changes and development of medical research writing. Atkinson’s corpora comprised samples of research articles from the Edinburgh Medical Journal between 1735 and 1985 at forty- to forty-five year intervals. Atkinson (1992: 343) used two types of text analysis, firstly, the MD method to perform a diachronic linguistic analysis of the articles, and secondly, a rhetorical, descriptive analysis of their broad genre characteristics. Atkinson (1992: 343-344) performed rhetorical and descriptive analysis of the corpora by examining the form and content of the texts and their formal and semantic characteristics, their degree of conventionalisation, relation of theory to data in each article examined and the stability of this relationship, design coherence and conventionalisation, and epistemological links between rhetorical forms and content. The results of Atkinson's (1992: 346-348) rhetorical analysis of the medical articles during the examined 250-year period revealed firstly, that, as for the contents of the articles, they changed from detailed descriptions of small numbers of medical cases towards more abstract and general reports of larger numbers of cases. Secondly, the style of discourse changed gradually from author prominent style to non-author-centred style. Thirdly, the articles' level of conceptual integration changed from fragmented coverage of unrelated cases towards more integrated and larger discussions and conventionalised experiment-based systems of text organisation.

In addition to the rhetorical and descriptive analysis of the broad genre characteristics of the articles, Atkinson used the MD method for their diachronic linguistic analysis. Atkinson's (1992: 346) corpora comprised ten texts per each EMJ's sampled forty to forty-five year period divided into five categories (case reports, disease reviews, treatment-focused reports, experimental reports, speeches, and category others), thus consisting of 70 texts and a total of 186,553 words. Atkinson's (1992: 344) main results suggest firstly, that the informational content of the medical research articles increased over the years (ref. Biber's Factor 1, informational vs. involved production), secondly, that the texts became less narrative (ref. Biber's Factor 2, narrative vs. non-narrative concerns*)*, and thirdly, that the level of explicit reference increased (ref. Biber's Factor 3, explicit vs. situation-dependent reference), and fourthly, that over the years the EMJ's articles have remained generally abstract and technical (ref. Biber's Factor 5, abstract vs. non-abstract information). In conclusion Atkinson (1992: 363) claims that medical research writing has changed gradually over the years, in contrast to the more general view of it having changed in more sudden steps resulting from changes in the scientific methods.

Atkinson (1996) used the MD method also in his study of scientific research writing in English from 1675 to 1975, for a corpus of articles from the Philosophical Transactions of the Royal Society of London, in addition to the rhetorical analysis of their broad genre characteristics. The results of Atkinson’s (1996: 357) MD analysis revealed that scientific research writing grew steadily more informational, there was a steady progression in a single direction from narrative to more non-narrative, a generally clear evolution toward a non-persuasive norm, development over time with early texts being abstract at a moderate level, but becoming extremely abstract in the modern period. Atkinson suggests (1996: 359-360) that the results of the study’s linguistic analysis, based on the MD method, and the rhetorical analysis of the articles’ genre characteristics, coincide on three major findings. Firstly, the linguistic shift from rather involved to very informational discourse (ref. Biber’s Factor 1) correlates with the rhetorical movement away from author-centred rhetoric. Secondly, in the development of a very abstract language (ref. Biber’s Factor 5) and object-centred rhetoric. Thirdly, in gradual loss of narrative elements over the years, which was shown both in the linguistic (ref. Biber’s Factor 2) and rhetorical levels.

3 THE PRESENT STUDY

Chapter 3 includes the research design of this study, with a presentation of the data and its analysis and the findings based on the factorial structure. It begins with section 3.1, where the a setting of the study is introduced with research questions.

3.1 Research design

For the purposes of this study the hypothesized text type setting of this study follows Widdowson's views on the description of scientific discourse as described above in section 2.1. Table 2 below presents a summary of the description of scientific discourse of software engineering with the hypothesised text types and their audiences in this study.

Table 2. Hypothesised text types and audiences

Language used in	**Text type**	**Audience**
scientific exposition	academic research articles (AR)	academic
scientific instruction	introductory textbooks (IT)	introductory university
scientific journalism	popular journals (PJ)	laypersons

The scientific language of software engineering is divided into three groups of text types: academic research articles (scientific exposition), introductory textbooks (scientific instruction), and popular journals (scientific journalism).

The audience is divided likewise into three groups. Firstly, the audience of academic research articles is called academic, because it concerns persons that have an academic education and who read and/or write research articles, and who are used to reading and even writing research articles. Secondly, students that read introductory textbooks about software engineering are at the first stages of their university studies and starting to get acquainted with and learn the scientific discourse of their discipline. Thirdly, the audience called layperson refers to ordinary people, who are non-scientists reading about matters such as computer programs and their applications, that is, software engineering, in popular journals such as PC Magazine and Macworld. The target audience of these magazines is not trained in software engineering matters or has a detailed knowledge of it.

Variation among these text types written for different audiences is studied in this study by the MD approach and the concept of dimension of linguistic variation. This method has been used in several studies on various ESP texts as described above, and in this study it is applied to texts of software engineering.

The following questions were chosen to guide the present study:

1) Can the hypothesised text types distinguished?
2) Do software engineering texts written for different audiences show variation?

To answer the first research question the co-occurrence relations among the linguistic features of sample texts are grouped into factors and mean factor scores are computed for them. The text types are arranged by their factor scores on graphic plots, which show the relations of the text types with one another on the extracted factors, that is, how different or similar, or distanced from each other, they are on each factor.

To answer the second research question the factor scores and the relations of the text types will be reviewed again when the the combinations of variables forming each factor are interpreted and labels are suggested as linguistic dimensions, based on the assessment of the communicative functions of the set of variables that load on each factor. The dimensions are given tentative labels that signify why the features co-occur in different text types.

This section gave an overview of the setting of this study with the research questions. In the next section I will present the data of this study. It includes the sample texts, the linguistic variables/features with descriptions and discussions of problems and decisions concerning the rules of coding, the coding and counting of the variables with a sample of coding, an overview of the statistical analysis, and the extracted factorial structure of four factors, which is the basis for the findings and results of this study.

3.2 Data and its analysis

For the three text types, academic research articles, introductory textbooks and popular journals, I randomly collected 8 sample texts of each from the field of software engineering. The corpus was collected at the library of the Department of Computer Science and Information Systems of the University of Jyväskylä. The research articles were taken from Management information systems research journals specialising in information systems and computer science, such as *MIS Quarterly Management Information Systems*, *Information and Management*, *IEEE Transactions on Software Engineering*, and *Information Systems Research*. Introductory textbooks were randomly selected from books with titles including preferably all of the terms *introduction*, *student*, and *software engineering*. The popular journal texts were taken from *PC Magazine* and *Macworld* by randomly selecting texts that concerned software, and not, for instance, computers or their presentations (hardware).

As a general guideline an attempt was made not to include texts published prior to the early 1980s, and, especially for longer research articles and textbooks, not to select samples only from the beginning or the end of the article in question. The view of Crookes (1986) was also taken into consideration, who, while selecting a corpus for his study, rejected articles whose main authors appeared to be non-natives of the English language.

The total corpus consisted of 24 texts, 8 texts of each text type academic research articles, introductory textbooks and popular journals, and their lengths ranged from 842 up to 1,098 words. The number of words of the academic research text type was 8,208, of introductory text books 8,008 and of popular journals 8,085 words, and thus the total of words in this study was 24,301. The mean length of the texts was 1,012 words. The complete list of the sample texts is below.

Sample texts

Academic research articles

Aiken, Milam, and Juditch Carlisle 1992. An automated idea consolidation tool for computer supported cooperative work, *Information and Management* 23, 373-382.
Extract from pages 375-377; 1034 words (file ARVAR18)

Davidson, Jack W., and Anne M. Holler 1992. Subprogram inlining: a study of its effects on program execution time, *IEEE Transactions on Software Engineering* 18, 89-102.
Extract from pages 91-92; 1033 words (file ARVAR20)

Eick, Stephen G., Joseph L. Steffen, and Eric E. Sumner, Jr. 1992. Seesoft - A tool for visualizing line oriented software statistics, *IEEE Transactions on Software Engineering* 18, 957-968.
Extract from pages 962-965; 1062 words (file ARVAR19)

Greif, Irene, Robert Seliger, and William Weihl 1992. A case study of CES: a distributed collaborative editing system implemented in Argus, *IEEE Transactions on Software Engineering* 18, 827-839.
Extract from pages 830-831; 1048 words (file ARVAR21)

Goodhue, Dalde L., Michael D. Wybo, and Laurie J. Kirsch 1992. The impact of data integration on the costs and benefits of information systems, *MIS QUARTERLY Management Information Systems* 16 (3), 293-311
Extract from pages 298-300; 1043 words (file ARVAR17)

Johnson, W. Lewis, Martin S. Feather, and David R. Harris 1992. Representation and presentation of requirements knowledge, *IEEE Transactions on Software Engineering* 18, 853-869.
Page 866; 842 words (file ARVAR23)

McLeod, Poppy Lauretta, and Jeffrey K. Liker 1992. Electronic meeting systems: evidence from a low structure environment, *Information Systems Research* 3, 195-223.
Extract from pages 217-220; 1098 words (file ARVAR24)

Taff, Louis M., James W. Borchering, and W. Richard Hudgins, Jr 1991. Estimeetings: development estimates and a front-end process for a large project, *IEEE Transactions on Software Engineering* 17, 839-849.
Extract from pages 841-842; 1048 words (file ARVAR22)

Introductory textbooks

Easteal, Charles, and Davies, Gordon 1989. *Software Engineering: analysis and design.* International software engineering series. London: McGraw-Hill.
Extract from pages 94-96; 1039 words (file ITVAR09)

Greenfield, Peter 1992. *Introduction to computing.* Cambridge: McGraw Hill.
Extract from pages 143-145, 871 words (file ITVAR10)
Extract from pages 172-174; 963 words (file ITVAR11)

Ince, D. C. 1989 (1990). *Software engineering.* London: Chapman and Hall.
Extract from pages 63-64; 1052 words (file ITVAR12)

Lamb, David Alex 1988. *Software engineering: planning for change.* Englewood Cliffs, N.J.: Prentice Hall.
Extract from pages 43-45; 1034 words (file ITVAR16)

Mynatt, Barbee Teasley 1990. *Software engineering with student project guidance.* Englewood Cliffs, NJ: Prentice Hall.
Extract from pages 192-194; 1035 words (file ITVAR15)

Shooman, Martin L. 1983 (1985). *Software engineering: design, reliability and management. International student edition.* Auckland: McGraw-Hill.
Extract from pages 38-40; 1021 words (file ITVAR13)

Sommerville, Ian 1985. *Software engineering.* 2nd edition. Wokingham: Addison-Wesley.
Extract from pages 93-96; 993 words (file ITVAR14)

Popular journals

Derfler, Frank J. Jr. 1992. TCP/IP packages for netware 3.11. Using the alphabet soup, *PC Magazine* 11 (13) 415-441
Extract from pages 420-425; 1023 words (file PJVAR03)

Miller, Michael J. 1992. Eight steps to easier software, *PC Magazine*, 11 (13) 81-82.
Extract from pages 81-82; 1002 words (file PJVAR07)

Miller, Michael J. 1993. Applications integration. Making your programs work together, *PC Magazine*, 12 (6) 108-137.
Extract from pages 109-112; 1033 words (file PJVAR02)

Raskin, Robin, and M. E. Kabay 1993. Antivirus software. Keeping your guard, *PC Magazine*, 12 (5) 209-269.
Extract from pages 211-218; 1033 words (file PJVAR05)

Roth, Steven 1993. Managing color, *Macworld*, January 1993, 148-155.
Extract from pages 153-155; 1079 words (file PJVAR06)

Seymour, Jim 1992. New interface dilemmas, *PC Magazine*, 11 (13) 99-100.
Extract from pages 99-100; 852 words (file PJVAR08)

Shipley, Chris 1993. Tracking dollars with sense, *PC Magazine*, 12, (1) 241-274.
Extract from pages 242-250; 1038 words (file PJVAR04)

Yakal, Kathy 1993. Personal tax preparation. Staying ahead of the tax man, *PC Magazine*, 12 (13) 225-252.
Extract from pages 226-230; 1025 words (file PJVAR01)

Selection of linguistic variables

The linguistic variables for this study were taken from the study carried out by Biber (1988). The number of variables had to be limited for the purposes of this study, and the decision was to include 40 %, that is, 27 pieces, of the 67 linguistic features of Biber's study. The features were selected from the six primary factors of Biber's (1988: 89-90) study, to have as representative coverage as possible. It was hypothesised that the more frequently occurring features of Biber's study would show variation also in a smaller corpus. The selection of the approximately 40 % of the linguistic features from each of the six primary factors was performed by starting from the highest values of both the positive and negative loading scores. The factor structure has linguistic features with positive loading scores on the top and features with negative scores on the bottom of the factors. These positive and negative loadings of the features occur in texts in a complementary pattern, that is, when a text has a high frequency of features with a positive loading, the features with a negative loading are non-existent in that text, and vice versa (Biber 1988: 88).

The linguistic variables of this study are listed below in Table 3, showing also the names of the factors in Biber's study (1988: 89-90), from which they were taken.

Table 3. Linguistic variables (adapted from Biber 1988: 89-90)

Factor 1, ´Involved vs. Informational Production´		
Features with positive loadings:	private verbs	[PRIVER]
	THAT deletion	[THADEL]
	contractions	[CONTRA]
	present tense verbs	[PRESTE]
	2nd person pronouns	[2NDPER]
	DO as pro-verb	[DOPROV]
	analytic negation	[ANANEG]
	demonstrative pronouns	[DEMPRO]

Features with negative loadings:	nouns	[NOUNS]
	word length	[WORDLE]
	prepositions	[PREPOS]
	type-token ratio	[TYPETO]
Factor 2, ´Narrative vs. Non-Narrative Concerns´		
	past tense verbs	[PASTTE]
	third person pronouns	[3RDPER]
	perfect aspect verbs	[PERFAS]
	(negative features in parentheses)	
Factor 3, ´Explicit vs. Situation-dependent Reference´		
	WH relative clauses on object positions	[WHREOB]
	pied piping constructions	[PIEDPI]
	WH relative clauses on subject positions	[WHRESU]
	---------------------------------	
	time adverbials	[TIMEAD]
	place adverbials	[PLACAD]
Factor 4, ´Overt Expression of Persuasion´		
	infinitives	[INFINI]
	prediction modals	[PREDMO]
	suasive verbs	[SUAVER]

	- - no negative features - -	
Factor 5, ´Abstract vs. Non-Abstract Information´		
	conjuncts	[CONJUN]
	agentless passives	[PASSIV]

	(one negative feature in parentheses)	
Factor 6, ´On-Line Informational Elaboration´		
	THAT clauses as verb complements	[THAVCO]
	demonstratives	[DEMONS]

The names in square brackets are the codes used for the features in this study. For instance, PRIVER is a code for private verbs. The mention of negative features in parentheses in Table 3 means that even though some features load on more than one factor, each feature is only included in the factor score of the factor on which it has the highest loading, whether it is a minus or plus value (Biber 1988: 93).

Descriptions of linguistics features with discussions

The 27 linguistic features mentioned above in Table 3 are listed below divided into grammatical categories according to Biber (1988: 221-245). Many of the features include short descriptions and discussion with references to Biber's studies (1986, 1988) and to the grammar of Quirk et. al. (1985). As the coding was not completely straightforward or unproblematic, they were consulted for support when encountering difficulties in the identification and interpretation of the linguistic features. The decisions concerning the rules of marking during coding that had to be made for the sake of consistency are also included.

A. Tense and aspect markers

1. past tense [PASTTE]

Biber sees (1988: 223) past tense forms as markers of narrative texts.

2. perfect aspect [PERFAS]

Quirk et al. (1985: 151) use the term *perfective* instead of *perfect* used by Biber (1988: 223), and according to them perfective aspect consists of the auxiliary HAVE + the *-ed* participle of a verb (for instance, *has examined*). According to Biber perfect aspect forms co-occur often with the past tense as markers of narrative texts (1988: 224). The count also includes the contracted forms of the auxiliary *have*.

3. present tense [PRESTE]

Biber (1988: 224) includes all base forms or third person singular present tense forms of verbs in the tagged dictionary in the identification of present tense forms, excluding infinitives. In this study present tense forms were counted likewise.

The base form is used, according to Quirk et al. (1985: 97), as a finite form, in addition to the present tense in all persons and numbers except the 3rd person singular, in the imperative and the present subjunctive. Therefore imperative and present subjunctive verbs were marked for the present tense, even though imperatives lack tense distinction (Quirk et. al 1985: 827) and the use of the present subjunctive relates more to mood than to tense (Quirk et al. 1985: 155).

An example of the present subjunctive:

(1) The integration procedure does require that some module *call* it and *provide* it ... (Text ITVAR16)

First person imperatives are also, according to Quirk et al. (1985: 829), formed by the verb *let* and followed by a subject in the objective case. In this study they were counted as present tense forms, of which here is an example:

(2) *Let* us now consider ... (Text ITVAR09)

Say as an explicit indicator of apposition (Quirk et al. 1985: 1307) was not marked for present tense, as in example:

(3) It may well be that when an organization comes to re-use some modules within a new system that it is to be implemented in Fortran 77, *say,* it finds that ... (Text ITVAR09)

That is is a conjunct (linguistic feature no. 20 CONJUN), and *is* was not marked in this connection for the present tense.

B. Place and time adverbials

4. place adverbials [PLACAD]

aboard, above, abroad, across, ahead, alongside, around, ashore, astern, away, behind, below, beneath, beside, downhill, downstairs, downstream, east, far, hereabouts, indoors, inland, inshore, inside, locally, near, nearby, north, nowhere, outdoors, outside, overboard, overland, overseas, south, underfoot, underground, underneath, uphill, upstairs, upstream, west

Biber (1988: 224) has taken the list above from Quirk et al. (1985: 516), and he has excluded items with other major functions, for example, *in*, *on*, which he says often mark logical relations in a text. This principle was followed also in this study.

5. time adverbials [TIMEAD]

afterwards, again, earlier, early, eventually, formerly, immediately, initially, instantly, late, lately, later, momentarily, now, nowadays, once, originally, presently, previously, recently, shortly, simultaneously, soon, subsequently, today, tomorrow, tonight, yesterday

Biber has taken this list from Quirk et al. (1985: 526ff), and he has excluded items with other major functions, for example, *last*, *next*, which according to him often mark logical relations in a text. This principle was followed also in this study.

One difficulty in the identification of time adverbials was that they are also used in other functions. For instance, *again* is also a formal reinforcing conjunct (Quirk et al. 1985: 635), and *now* is also a resultive conjunct (Quirk et al. 1985: 635). Examples of these:

(4) *Again*, some let you compare your major income and expense categories with U.S. averages. (Text PJVAR01)

(5) *Now*, with more sophisticated applications being introduced for Windows ... (Text PJVAR02)

Today is not always a time adverbial. According to Quirk et al. (1985: 478, 501-502) adverbs may operate grammatically as part of phrases, which as a whole realize other elements. In the following example, *today* was part of the subject noun phrase, and it was not counted as a time adverbial:

(6) ..., *today*'s better antivirus software products... (Text PJVAR05)

According to Quirk et al. (1985: 503), the grammatical function of *today* may also be ambiguous, as it may be part of a subject noun phrase, as, for instance, in the sentence “The pace of life *today* is proving too fast”. However, Quirk et al. (1985: 503) state that the position does not exclude the occurrence of the adverbial function, as in sentence “Mr Jones *today* must be heartbroken”. Accordingly, *today* was marked as a time adverbial in the following example:

(7) About half of the products that are on the market *today* offer an interview option ... (Text PJVAR01)

According to Biber (1988: 224), place and time adverbials refer to the physical and temporal context of the text. They are also suggested to mark more concrete, situated contents by reference to external situations (Biber 1986: 396).

C. Pronouns and pro-verbs

6. second person pronouns [2NDPER]

you, your, yourself, yourselves (plus contracted forms)

7. third person pronouns (excluding *it*) [3RDPER]

she, he, they, her, him, them, his, their, himself, herself, themselves (plus contracted forms)

According to Biber (1988: 225, 396), third person personal pronouns mark relatively inexact reference to persons outside of the immediate interaction and they co-occur frequently with past tense and perfect aspect forms referring to a removed, narrative context, instead of immediate reference indicated by the present tense and adjectives.

8. demonstrative pronouns [DEMPRO]

(*that, this, these, those* as pronouns; e.g., *this is ridiculous*)

Demonstrative pronouns need to be separated from relative pronouns and complementizers, etc. (Biber 1988: 226).

According to Biber (1988: 226), demonstrative pronouns can refer to an entity outside the text or to a previous referent in the text itself. When referring to a previous referent in the text in question, it can refer to a specific nominal entity or to an inexplicit, often abstract, concept (e.g., *this shows* ...).

9. pro-verb *do* [DOPROV]

(e.g., *the cat did it*)

Do is a pro-verb, a substitute, as a main verb, and in this function it must be distinguished from *do* as an auxiliary verb (Quirk et al. 1985: 874). In this study the cases of *do* as a pro-verb come mostly from transitive cases, where *do* combines with a pronoun object to act as a propredication referring to some unspecified action or actions (Quirk et al. 1985: 134-135). The pronoun object may be personal (*it*), demonstrative (*this/that*), interrogative (*what*), or indefinite (*nothing/anything*, etc).

Do is also used in informal discourse as a general-purpose agentive transitive verb and a so called get-passive (Quirk et al. 1985: 135, 160-161). These are not cases of pro-verb *do* and not counted. Examples:

(8) When you *do* your taxes manually... (Text PJVAR01)
(9) … get the job *done* … (Text ARVAR17)

10. Total other nouns [NOUNS]

In this study the counting of nouns was made in accordance with the feature which in Biber's (1988: 228) study was called *total other nouns.* Nominalizations, that is, words ending in *-tion*, *-ment*, *-ness* and *-ity* or their plural forms, and gerunds, participle forms serving nominal functions, were excluded.

Help for the identification of total other nouns and their separation from gerunds was found in Quirk et al. (1985: 1521), who use the term verbal noun for the 'gerund' class of nouns ending in *-ing*.

As the tagged dictionary Biber used in his study is not available, some rules concerning the counting of proper nouns had to be made. Proper nouns were not counted, for instance, as names of software (in text PJVAR4 *CheckFree* 3.0, *Cheque-It-out* 2.0), except when they were common nouns (for instance, in PJVAR4 Microsoft *Money*, 2.0 Quicken for *Windows* 2.0). Names of persons were not counted (for instance, *Patricia Hoffman* in text PJVAR05), or abbreviations, because the tagged dictionary used in Biber's study was not available for checking, and the use of abbreviations in the texts under study was abundant. One could not be sure what abbreviations are old and common knowledge, being most probably included in dictionaries (for instance, *PC, DOS, Unix, RAM, ASCII*), and which are newer and may not be included in dictionaries as such (for instance *LAN, GUI, OLE, Shift-Del, Ctrl-Dels, Ctrl-V*). Due to this decision also such general abbreviations as *U.S., Sec., Fig*. were not marked for nouns.

Conversions from phrases or affixes to nouns, such as, for instance, *Shift-Ins, pop ups,* and *drag-and-drop* in text PJVAR08 were not marked as total other nouns. Nouns consisting of two words divided by a slash were also not counted, as, for instance, *input/output* in text ITVAR10.

Not all supposed nouns were counted, as, for instance, in a case of a printing error such as:

(10) Computer hardware functional *unis* [sic] may use .... (Text TVAR11)

If a text has a lot of nominal content, it is suggested to indicate a high informational focus, instead of interpersonal or narrative foci (Biber 1988: 227).

E. Passives

11. agentless passives [PASSIV]

Agentless passives were counted on the basis of the participle forms. For instance, the following example includes two passive forms that were marked and counted:

> (11) ... form that can be *discussed* and *validated* ... (Text ITVAR11)

According to Biber (1988: 228) passives are one of the most important surface markers of the decontextualized or detached style of writing, and when the agent is dropped altogether, as it is counted in this study, it results in a static, abstract presentation of information.

F. Subordination features

Complementation

12. *that* clauses as verb complements [THAVCO]

(e.g., *I said that he went*)

> (12) We referred to some research that showed that a programmer's comprehension ... (Text PJVAR07)

According to Biber (1986: 394-395) *that*-clauses co-occur frequently with many features that can mark interpersonal interaction and personal involvement, such as first and second person pronouns and subordinating clauses, and according to him the extensive use of subordination is associated with production constraints characteristic of speech.

13. *to*-infinitives [INFINI]

The algorithm Biber (1988: 232) uses in his study to count infinitives groups together all infinitival forms: complements to nouns, adjectives, and verbs, as well as adverbial purpose clauses.

According to Biber (1988: 232) the distribution and discourse functions of infinitives seem to be less marked than those of other types of subordination.

Relatives

14. WH relative clauses on subject position [WHRESU]

(e.g., *the man who likes popcorn*)

WH refers to pronouns *who, whom, whose, which* (Biber 1988: 223). Biber excludes indirect WH questions with verbs *ask* and *tell*, e.g. *Tom asked the man who went to the store* (Biber 1988: 235).

15. WH relative clauses on object positions [WHREOB]

(e.g., *the man who Sally likes*)

WH refers to pronouns *who, whom, whose, which* (Biber 1988: 223). Biber excludes indirect WH questions with verbs *ask* and *tell* (Biber 1988: 235).

16. pied piping constructions [PIEDPI]

(e.g., *the manner in which he was told*)

Relative clause structures, which have *wh*-pronouns with initial prepositions are called by Biber (1988: 235) pied piping constructions.

Wh-pronouns are used predominantly in formal English, and initial prepositions are normally avoided in more informal use (Quirk et al. 1985: 1252-1253).

G. Prepositional phrases

17. total prepositional phrases [PREPOS]

against, amid, amidst, among, amongst, at, besides, between, by, despite, during, except, for, from, in, into, minus, notwithstanding, of, off, on, onto, opposite, out, per, plus, pro, re, than, through, throughout, thru, to, toward, towards, upon, versus, via, with, within, without

Biber (1988: 237) has taken this list of prepositions from Quirk et al. (1985: 665-667), and he has excluded lexical items that have some other primary function, such as place or time adverbial, conjunct, or subordinator (e.g., *down*, *after*, *as*).

Prepositions were difficult to mark, because they are often formally identical with and semantically similar to adverbs (Quirk et al. 1985: 662). There are also many multi-word verbs, where the lexical verb is followed by prepositions and adverbs that are morphologically invariable (Quirk et al. 1985: 1150-1151). Such verbs are called prepositional and phrasal verbs, respectively. An example of a transitive phrasal verb in this study in which *out* is an adverb particle (Quirk et al. 1985: 1177), and not counted in total prepositional phrases:

(13) ... every time you *fill out* a supporting worksheet ... (Text PJVAR01)

According to Biber (1988: 237), prepositions are an important device for packing high amounts of information into academic nominal discourse. Biber (1986: 395, 404) also claims that they tend to co-occur for instance with nominalizations, passives, and conjuncts, sharing a function which marks a highly abstract, nominal content and a highly learned style such as academic prose, official documents, and professional letters.

H. Lexical specificity

18. type/token ratio [TYPETO]

The type/token ratio refers to the number of different lexical items in a text, as a percentage (Biber 1988: 238). In the present study the type/token ratio is computed in the same way as in Biber's study (1988: 239) by counting the number of different lexical items that occur in the first 400 words of each text, and then dividing it by four. The counting was performed by the Computing Centre of the University of Jyväskylä.

19. mean word length [WORDLE]

The mean length of the words in a text, in orthographic letters. The counting was performed by the Computing Centre of the University of Jyväskylä.

I Lexical classes

20. **conjuncts** (e.g., *consequently, furthermore, however*) **[CONJUN]**

alternatively, altogether, consequently, conversely, eg, e.g., else, furthermore, hence, however, i.e., instead, likewise, moreover, namely, nevertherless, nonetheless, notwithstanding, otherwise, rather, similarly, that is, therefore, thus, viz.; in + comparison / contrast / particular / addition / conclusion / consequence / sum / summary

/ any event / any case / other words; for + example / instance; by + contrast/comparison; as a + result / consequence; on the + contrary / other hand

Conjuncts are suggested to mark logical relations between clauses (Biber 1988: 239). They are also claimed to co-occur for instance with nominalizations, prepositions, and passives, sharing a function, which marks an abstract and learned style (Biber 1986: 395).

21. demonstratives [DEMONS]

that/this/these/those

Demonstratives are excluded from demonstrative pronouns (feature no. 8 DEMPRO), as well as *that* as relative, complementizer, or subordinator (Biber 1988: 241) .

J. Modals

22. prediction modals [PREDMO]

will/would/shall (+ contractions)

K. Specialized verb classes

Private and suasive verbs are included in this study in the specialized verb classes. According to Biber (1988: 242), restricted classes of verbs have specific functions, for instance, private verbs express intellectual states *(believe*) or non-observable intellectual acts (*discover*), and suasive verbs imply intentions to bring about some change in the future (*command, stipulate*).

When listing these verbs in his study, Biber (1988: 242) refers to Quirk et al. (1985: 1181-1183), and gives some examples. Some verbs belong to more than one group, and as Biber does not give any indication of how to treat these verbs, they were excluded in this study. Therefore the private and suasive verbs mentioned by Quirk et al. (1985: 1181-1182) were included in the count, excluding such verbs that belong both to the private and the suasive class of verbs (*decide, determine, ensure*), as well as verbs that belong both to the suasive and the public verb group (*agree, concede, insist, pronounce, suggest*). Biber (1988: 131) also counts some verbs that are not mentioned by Quirk et al., for instance, the private verb *love*, which is mentioned in his sample text for Dimension 1.

According to Biber (1988: 242), all present and past tense forms of these specialized verbs were counted, including infinitives, so also in this study all finite and nonfinite forms of the above mentioned verbs were counted.

23. private verbs [PRIVER]

accept, anticipate, ascertain, assume, believe, calculate, check, conclude, conjecture, consider, decide, deduce, deem, demonstrate, determine, discern, discover, doubt, dream, ensure, establish, estimate, expect, fancy, fear, feel, find, foresee, forget, gather, guess, hear, hold, hope, imagine, imply, indicate, infer, insure, judge, know, learn, mean, note, notice, observe, perceive, presume, presuppose, pretend, prove, realize, reason, recall, reckon, recognize, reflect, remember, reveal, see, sense, show, signify, suppose, suspect, think, understand

24. suasive verbs [SUAVER]

agree, allow, arrange, ask, beg, command, concede, decide, decree, demand, desire, determine, enjoin, ensure, entreat, grant, insist, instruct, intend, move, ordain, order, pledge, pray, prefer, pronounce, propose, recommend, request, require, resolve, rule, stipulate, suggest, urge, vote

L. Reduced forms

25. contractions [CONTRA]

Major types of contractions in writing, not corresponding to reductions in speech, are, according to Quirk et al., the following:

- forms of BE *‘m, aren’t, ‘s, ‘re, isn’t, wasn’t, weren’t* (1985: 130)
- forms of HAVE: *‘ve, ‘s, haven’t, hasn’t, ‘d, hadn’t* (1985: 131)
- forms of DO: *don’t, doesn’t, didn’t* (1985: 133)
- forms of modal auxiliaries: *can’t, couldn’t, mayn’t, mightn’t, shan’t, shouldn’t, ‘ll, won’t, ‘d, wouldn’t, mustn’t* (1985: 135)
- personal pronoun contraction of *us* in *let’s* (1985: 148)

All contractions were counted in this study. They are dispreferred in formal, edited writing (Biber 1988: 243)

26. **subordinator *that* deletion** (e.g., *I think* [*that*] *he went*) **[THADEL]**

According to Quirk et al. (1985: 1049) nominal *that*-clauses may function, for instance, as direct object, and when the *that*-clause is direct object or complement, the conjunction *that* is frequently omitted except in formal use (I know *it's late*). There are very few deletions of subordinator-*that* in edited writing (Biber 1988: 244).

M. Negation

27. **analytic negation** (e.g., *that's not likely*) **[ANANEG]**
(also contracted forms)

Biber counts their frequencies of synthetic and analytic negation as separate features. This study includes only the not-negation, which is suggested to be more colloquial and more fragmented than synthetic negation (Tottie 1983, as quoted by Biber 1988: 245).

Coding and counting of variables

I called the three groups of eight texts as PJVAR01…08, ITVAR09…16, and ARVAR17…24, referring to the VARiables of the text types of Popular Journals, Introductory Textbooks, and Academic Research articles. The linguistic features were given short names of six letters, or numbers in the case of 2nd and 3rd person personal pronouns, to give some "sound effect" to make their manual counting easier and to avoid spelling errors. The corpus texts were scanned to file format, to make their processing easier by the WordPerfect word processing program.

The linguistic variables were marked in the text files with the Edit commands Copy and Paste, and then the markings were counted by using the Find command.

An example of the coding of linguistic features in this study is presented below in Table 4.

Table 4. Example of coding (file ARVAR17)

> Recent additions to [PREPOS] organizational information processing theory [NOUNS] take [PRESTE] the concept [NOUNS] of [PREPOS] uncertainty [NOUNS] as used by [PREPOS] Galbraith (1973) and Tushman and Nadler (1978) and break [PRESTE] it into [PREPOS] information requirements of [PREPOS] two types [NOUNS]: uncertainty [NOUNS] and equivocality (Daft and Lengel, 1986). This [DEMONS] distinction can help us understand [PRIVER] when computerized information systems [NOUNS] (whether with [PREPOS] data [NOUNS] integration or not [ANANEG]) may be of [PREPOS] little assistance [NOUNS] in [PREPOS] meeting a firm's [NOUNS] most pressing information needs [NOUNS]. *Uncertainty* [NOUNS] by [PREPOS] this [DEMONS] new definition is [PRESTE] the absence [NOUNS] of [PREPOS] specific, needed information. For example [CONJUN], a manager [NOUNS] might want to [INFINI] know [PRIVER] whether sales [NOUNS] of [PREPOS] widgets [NOUNS] dropped [PASTTE] in [PREPOS] the southern region [NOUNS] last month [NOUNS]. The question is [PRESTE] clear, and given data [NOUNS] on [PREPOS] specific variables [NOUNS], the uncertainty [NOUNS] would [PREDMO] be removed [PASSIV].
>
> *Equivocality* means [PRESTE] [PRIVER] [THADEL] there are [PRESTE] multiple, conflicting interpretations of [PREPOS] a situation. When equivocality is [PRESTE] present, the precise information needed to [INFINI] resolve [SUAVER] the situation is [PRESTE] not [ANANEG] clear. For example [CONJUN], the same manager [NOUNS] might want to [INFINI] know [PRIVER] *why* sales [NOUNS] dropped [PASTTE] in [PREPOS] the southern region [NOUNS] last month [NOUNS]. For [PREPOS] such a question, a number [NOUNS] of [PREPOS] types [NOUNS] of [PREPOS] information might be appropriate, including the attitude [NOUNS] and behavior [NOUNS] of [PREPOS] the sales [NOUNS] force [NOUNS], the actions of [PREPOS] competitors [NOUNS], the economic situation, trends [NOUNS] in [PREPOS] customer [NOUNS] satisfaction, the weather [NOUNS], etc. Competing perspectives [NOUNS] might be as useful to [PREPOS] the manager [NOUNS] as factual data [NOUNS].
>
> In [PREPOS] general, uncertainty [NOUNS] can be reduced by [PREPOS] a sufficient *amount* [NOUNS] of [PREPOS] information, while equivocality can be reduced by [PREPOS] sufficiently *rich* information (Daft and Lengel, 1986). Certain types [NOUNS] of [PREPOS] information processing mechanisms [NOUNS] (such as computerized information systems [NOUNS] and, by [PREPOS] implication, systems [NOUNS] with [PREPOS] integrated data [NOUNS]) provide [PRESTE] large amounts [NOUNS] of [PREPOS] information and can thus [CONJUN] help reduce uncertainty [NOUNS]. However [CONJUN] they [3RDPER] are [PRESTE] not [ANANEG] as rich a source [NOUNS] as other information processing mechanisms [NOUNS] (such as face-to-face meetings [NOUNS]) and thus [CONJUN] are [PRESTE] not [ANANEG] as effective in [PREPOS] reducing equivocality.

Statistical analysis

After the coding I filled out the frequencies of the linguistic features in each sample text on an Excel file, for the statistical processing in the Computing Centre of the University of Jyväskylä. The linguistic features no. 18, type/token ratio and no. 19, mean word length, were also counted there.

Like in Biber’s study (1988: 75-76), the frequency counts of the linguistic features were normalised to a text length of 1,000 words, to have comparable frequencies of occurrence and to enable direct comparisons of the frequencies in texts, despite varying text lengths. The frequencies were standardised to a mean of 0.0 and standard deviation of 1.0, to have the values of the features comparable because they are translated to a single scale; see Biber (1988: 94) for more details.

The extraction method used in the factor analysis of this study was Principal Axis Factoring. Factor analysis combines variables that co-occur in the texts, reducing the number of variables to a small number of uncorrelated factors. A rotation process is required to receive a simple structure in which each linguistic feature loads on as few factors as possible. Without the rotation process the first factor would account for the greatest proportion of the variance and most of the features would load on this factor. The rotation method used for this study was Promax with Kaiser Normalization. Further details on factor analysis methodology are available in Biber’s study (1988: 79-97).

The following linguistic features were left out of the analysis in the statistical run due to their lower frequencies of occurrence: time adverbials (No. 5 TIMEAD), WH relative clauses on object positions (No. 15 WHREOB), pied piping constructions (No. 16 PIEDPI), and suasive verbs (No. 24 SUAVER). *To*-infinitives (No. 13 INFINI) and prediction modals (No. 17 PREDMO) dropped out in a later stage.

Factorial structure

In this study four factors were extracted and rotated. Table 5 represents the salient positive and negative loadings on each of the four factors of this analysis. It is based on the analysis of 27 linguistic features counted in 28 texts.

Table 5. Summary of the factorial structure

Factor 1		**Factor 2**		**Factor 3**		**Factor 4**	
CONTRA	0.87	THAVCO	0.89	PRESTE	0.69	WORDLE	0.73
2NDPER	0.85	THADEL	0.72	PLACAD	0.48	TYPETO	0.70
3RDPER	0.49	PRIVER	0.45	DOPROV	0.45	DEMONS	0.68
ANANEG	0.48	- - - - - - - - - - - - -		(CONJUN	0.36)	DEMPRO	0.55
(TYPETO	0.39)	NOUNS	- 0.53	- - - - - - - - - - - - -		(WHRESU	0.35)
- - - - - - - - - - - - - - -				PASTTE	- 0.70	PERFAS	0.31
PASSIV	- 0.67			(PREPOS	- 0.49	- - - - - - - - - - - - -	
PREPOS	- 0.69			(PRIVER	- 0.39)	(NOUNS	- 0.41)
CONJUN	- 0.65						
WHRESU	- 0.61						
(WORDLE	- 0.31)						

The linguistic features with positive loading scores are on the top and the features with negative scores on the bottom of factors. The positive and negative loadings of the features on a factor occur in texts in a complimentary manner. When a text includes a high frequency of features with a positive loading, the group of features with a negative loading is infrequent or non-existent in that text, and vice versa. In the interpretation of the factors both the positive and negative groups of features are taken into consideration (Biber 1988: 88).

Each feature is included only once in the factor score of the factor on which it has the highest loading, whether it is a minus or plus value (Biber 1988: 93). That is why, for instance, the smaller loadingsof type/token ratio (TYPETO) 0.39 and word length (WORDLE) – 0.31 on Factor 1, marked in parenthesis, are not included in the counting of the factor scores of that factor, but their higher respective loadings of 0.70, and 0.73 are included in the factor scores of Factor 4.

This section presented the data of this study, with a list of sample texts and a detailed description of the linguistic variables with discussions about the problems and decisions concerning their identification and marking. The coding and counting of variables was reported with a coding sample, and an overview given of the statistical analysis and the extraction of four linguistic factors, concluding with the final factorial structure of four factors.

The following three sections will present the findings based on the frequencies of the linguistic features and the factorial structure.

3.3 Findings

The purpose of this study is to find out if there is variation in the scientific text types of software engineering written for different audiences. The sample texts are divided in three text types; academic research articles, introductory textbooks and popular journals. The audience is divided likewise into three groups; academic, introductory, and laypersons. The method is provided by the MD method of linguistic analysis and the concept of dimension of linguistic variation. The co-occurrence relations of the linguistic features of sample texts are grouped into factors, which are then interpreted as linguistic dimensions, based on the assessment of the communicative functions of the features. The dimensions are given tentative labels to suggest why the features co-occur in different text types.

In the next section the frequencies of the linguistic features in the corpus of this study will be reported, with a review of their similarities and differences.

3.3.1 Frequencies of linguistic features

This section presents the frequencies of the linguistic features in the corpus of this study and their statistics. The analys of the frequencies will give the first insight of the processing of the data to find out if variations will show up.

The frequencies of the linguistic features in each text and text type with their mean values, normalised to text lengths of 1,000 words, are presented below in Table 6.

Table 6. Frequencies of linguistic features

Frequencies of linguistic	PASTTE	PERFAS	PRESTE	PLACAD	TIMEAD	2NDPER	3RDPER	DEMPRO	DOPROV	NOUNS	PASSIV	THAVCO	INFINI	WHRESU	WHREOB	PIEDPI	PREPOS	TYPETO	WORDLE	CONJUN	DEMONS	PREDMO	PRIVER	SUAVER	CONTRA	THADEL	ANANEG
features / text	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.	21.	22.	23.	24.	25.	26.	27.
PJVAR01	8	4	75	1	1	43	11	3	1	230	7	0	14	0	0	0	92	57	5	2	10	4	5	7	12	0	7
PJVAR02	11	2	78	2	4	28	7	9	0	222	7	1	13	0	0	1	114	54	5	4	8	3	5	3	6	0	7
PJVAR03	28	0	42	0	2	0	5	0	0	242	1	5	21	1	0	0	122	51	5	5	14	1	12	4	0	0	2
PJVAR04	15	3	61	1	4	40	16	1	1	239	2	1	24	0	0	1	90	58	5	5	9	7	13	0	12	0	1
PJVAR05	27	1	49	1	0	10	12	3	0	209	12	2	27	0	0	2	103	54	5	3	7	4	8	1	3	0	4
PJVAR06	6	2	62	2	3	11	5	6	1	233	4	1	19	1	0	0	112	55	5	3	8	15	6	3	5	0	2
PJVAR07	1	4	81	0	1	26	10	7	0	206	0	1	31	0	0	1	69	50	5	1	7	9	10	4	26	0	19
PJVAR08	14	2	53	0	0	11	2	5	0	164	4	2	15	0	0	0	76	55	5	2	13	7	14	0	16	0	10
Total	**110**	**18**	**501**	**7**	**15**	**169**	**68**	**34**	**3**	**1745**	**37**	**13**	**164**	**2**	**0**	**5**	**778**	**-**	**-**	**25**	**76**	**50**	**73**	**22**	**80**	**0**	**52**
Mean	**13.6**	**2.2**	**62.0**	**0.9**	**1.9**	**20.9**	**8.4**	**4.2**	**0.4**	**215.8**	**4.6**	**1.6**	**20.3**	**0.2**	**0.0**	**0.6**	**96.2**	**54.0**	**5.0**	**3.1**	**9.4**	**6.2**	**9.0**	**2.7**	**9.9**	**0.0**	**6.4**
ITVAR09	25	3	67	1	9	7	4	7	1	179	30	9	16	2	0	1	102	55	5	8	4	10	18	3	0	0	8
ITVAR10	0	4	57	3	2	0	10	3	4	174	24	3	20	2	0	0	69	43	5	10	7	2	8	3	0	1	2
ITVAR11	0	2	59	0	5	0	4	2	0	220	29	1	12	2	0	0	117	38	5	12	7	2	6	3	0	0	2
ITVAR12	0	1	70	1	2	1	1	3	0	259	30	0	7	6	0	0	123	39	5	10	10	10	1	3	0	0	2
ITVAR13	1	4	76	0	2	0	2	5	1	240	37	3	22	9	0	0	123	53	5	8	14	7	8	2	0	0	6
ITVAR14	3	1	61	2	3	0	4	8	0	195	20	3	20	4	0	1	83	46	6	12	18	3	5	3	0	0	5
ITVAR15	2	8	54	0	1	3	7	4	1	231	34	1	25	0	0	0	109	42	5	9	5	6	12	4	0	0	3
ITVAR16	1	0	85	0	0	7	2	5	1	199	3	5	22	0	0	0	115	39	5	5	6	3	12	4	0	1	10
Total	**32**	**23**	**529**	**7**	**24**	**18**	**34**	**37**	**8**	**1697**	**207**	**25**	**144**	**25**	**0**	**2**	**841**	**-**	**-**	**74**	**71**	**43**	**70**	**25**	**0**	**2**	**38**
Mean	**4.0**	**2.9**	**66.1**	**0.9**	**3.0**	**2.2**	**4.2**	**4.6**	**1.0**	**211.9**	**25.8**	**3.1**	**18.0**	**3.1**	**0.0**	**0.2**	**105.0**	**44.3**	**5.1**	**9.2**	**8.9**	**5.4**	**8.7**	**3.1**	**0.0**	**0.2**	**4.7**
ARVAR17	4	0	42	0	1	0	5	4	0	210	7	5	14	0	0	1	125	46	6	12	7	9	9	2	0	1	9
ARVAR18	1	8	65	0	0	0	5	1	0	273	25	1	19	8	0	3	123	54	6	12	12	5	9	1	0	0	5
ARVAR19	11	6	55	0	4	0	12	3	0	255	27	5	29	1	0	0	101	48	5	3	11	5	10	11	0	0	0
ARVAR20	20	0	28	0	1	0	2	1	0	288	22	1	12	0	0	2	156	46	5	2	6	2	11	1	0	0	4
ARVAR21	4	1	74	2	0	0	1	1	0	203	27	4	27	0	0	0	116	42	5	14	2	1	5	2	0	0	8
ARVAR22	3	6	72	2	5	0	1	5	0	238	23	0	9	3	0	5	133	50	5	4	12	5	9	6	1	0	4
ARVAR23	4	4	49	0	2	0	6	9	0	123	10	0	30	2	0	1	96	52	6	5	14	5	7	2	0	0	2
ARVAR24	22	8	43	0	1	0	10	4	0	196	8	12	12	1	0	2	121	53	5	10	21	11	15	3	0	2	7
Total	**69**	**33**	**428**	**4**	**14**	**0**	**42**	**28**	**0**	**1786**	**149**	**28**	**152**	**15**	**0**	**14**	**971**	**-**	**-**	**62**	**85**	**43**	**75**	**28**	**1**	**3**	**39**
Mean	**8.4**	**4.0**	**52.1**	**0.5**	**1.7**	**0.0**	**5.1**	**3.4**	**0.0**	**217.6**	**18.2**	**3.4**	**18.5**	**1.8**	**0.0**	**1.7**	**118.3**	**49.0**	**5.4**	**7.6**	**10.4**	**5.2**	**9.1**	**3.4**	**0.1**	**0.4**	**4.8**

Linguistic features, variables of this study, are listed on the top of Table 6 above, numbered from 1 to 27, and the texts of the three text types are listed on the left-hand column from top to bottom. The texts of the public journals text type are numbered from PJVAR01 to PJVAR08, and likewise the texts of introductory textbooks from ITVAR09 to ITVAR16, and of academic research articles from ARVAR17 to ARVAR24. Table 6 shows the frequencies of the linguistic features in each sample text and text type, also with their total frequencies and mean values.

Before making any comparisons it is worth remembering that the frequencies of various features are not directly comparable with each other, because, for instance, the linguistic feature no. 18, TYPETO, refers to the type/token ratio indicated in percentage, and no. 19, WORDLE, is the mean length of the words in a text in orthographic letters.

The frequencies of the linguistic features in Table 6 shows that some features occur with similar frequencies and are rather evenly distributed in the three text types, whereas some of them have rather different frequencies. Some of these similarities and differences are reviewed below with examples.

There are three features that are almost the same in all text types. Firstly, the mean word length, feature no. 19 WORDLE, is 5 letters in all popular journals texts, and either 5 or 6 letters in the texts of introductory textbooks and academic research articles.

Secondly, all three text types have similar nominal contents. The linguistic feature of total other nouns, no. 10 NOUNS, occurs with almost identical frequencies; the mean in PJ texts is 216, in IT texts 212, and in AR texts 218. Thirdly, feature no. 23 PRIVER, has almost similar frequencies in all text types. The mean frequency of private verbs is 9.0 for PJ texts, 8.7 for IT texts and 9.1 for AR texts.

There are two features that, in addition to having similar frequencies, also have small frequencies in common. Subordinator *that* deletion, feature no. 26 THADEL, and *that* clauses as verb complements, no. 12 THAVCO, have small and similar frequencies in all three text types. There are no occurrences of subordinator *that* deletion in PJ texts, whereas in IT texts the mean frequency is 0.2, and in AR texts 0.4. Subordinator *that* deletion has a mean frequency of 1.6 in PJ texts, 3.1 in IT texts, and 3.4 in AR texts.

Next we have a look at some features that are distributed unevenly, with differing frequencies, in the texts. Firstly, in the text type of public journals the feature no. 1 PASTTE, past tense, has the highest mean frequency of 13.6 per text with 110 occurrences, whereas in academic research articles the mean is 8.4, with 69 past tense forms, and in texts of introductory textbooks the mean frequcncy is 32 occurrences, with a mean of 4.0.

Secondly, the second and third person pronouns occur more frequently in public journals texts than in the other text types. Second person pronouns occur in PJ texts with a highest mean frequency of 20.9, whereas in introductory textbooks the mean is 2.2, and in academic research articles there are no second person pronouns at all. As for third person pronouns the difference between the text types is smaller. The mean frequency of third person pronouns in PJ texts is 8.4, in IT texts 4.2, and in AR texts 5.1.

Thirdly, agentless passives, feature no. 11 PASSIVE, occur most frequently in the texts of introductory textbooks, with a mean frequency of 25.8. In the text type of popular journals the mean value of agentless passives is lowest, 4.6.

Fourthly, contractions, no. 25 CONTRA, occur in public journals texts most frequenctly, that is, 80 times, with a mean of 10 per text, whereas it is non-existent in the texts of introductory textbooks and in the texts of academic research articles there is only one contraction.

Based on the above reported differences in the frequencies of the linguistic features between the three text types we can expect the text types to be distinguished from one another to some extent and variation between them to show up. The disctinction and variation may mostly result from differing frequencies of features such as, for instance, past tense forms, second and third person pronouns, agentless passive, and contractions.

We will also find out what significance the features that have similar and even rather low frequencies will have when the variation between the text types is studied further.

As the MD method is based on the co-occurrence of various linguistic features in texts and the statistics do not give direct answer to the research questions, we continue in the next section to see the relations of the three text types regarding each factor.

The statistics will be used in the following sections to support the findings to answer the research questions.

3.3.2 Distinguishing of text types

In this section I report the findings of the MD analysis to the first research question *Can the hypothesised text types be distinguished*?

To answer the first research question we need to analyse the mean factor scores, also called factor scales that show the relations of the text types with one another on the four factors. The text types are arranged by their factor scores in relation to one another on graphic plots. The factor scores are computed by summing for each text the frequency of both the positive and negative loadings on each factor and averaging the final score for each text type (Biber 1988: 93). The computation was performed at the Computing Centre of the University of Jyväskylä. As mentioned above, each feature is included in the computation of only one factor score. Below in section 3.3.3 the factor scores and the relations of the text types will be reconsidered again when the combinations of variables that form each factor are assessed and labels are suggested for linguistic dimensions by the set of variables that load on each factor.

Figure 4 below shows the relations of the text types of this study in relation to one another, based on their factor scores, with respect to Factors 1, 2, 3 and 4.

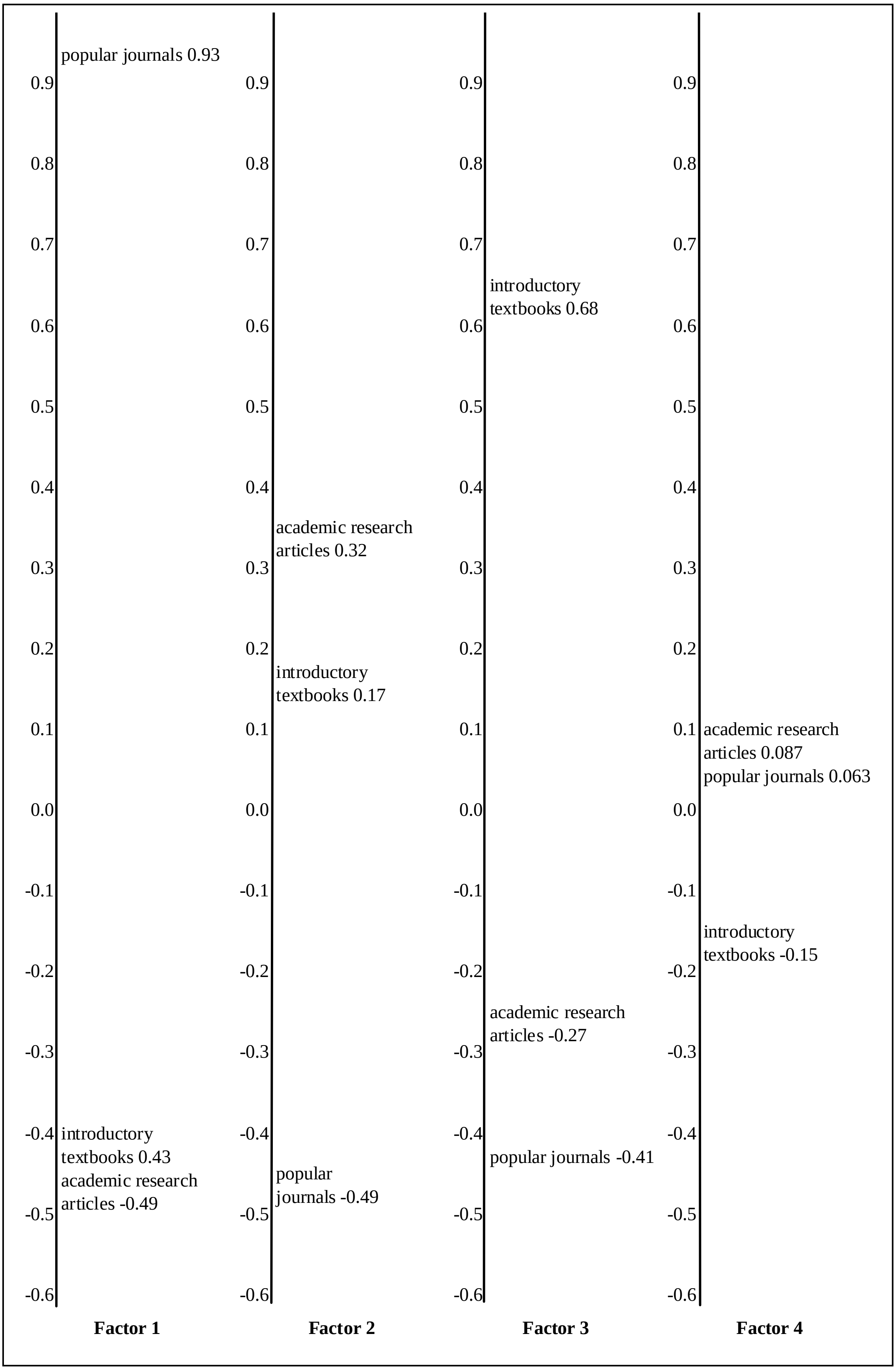


Figure 4. Text types in relation to one another with respect to Factors 1, 2, 3 and 4

In Figure 4 the three text types are arranged on Factors 1-4 in relation to each other, based on their factor scores. Regarding these factors and their linguistic features the text types are to some extent distinguished. On the one hand, the text type of public journals has the highest factor score on Factor 1 and a position distanced from the other text types, whereas the text types of introductory textbooks and academic research articles have lower, almost similar scores. On the other hand, on Factor 2, the arrangement of the text types is similar, but the other way round. Popular journals texts have the lowest score, whereas academic research articles and introductory textbooks have higher scores and they are closer to one another.

The plots of Factors 1 and 2 suggest that regarding these factors the text types of introductory textbooks and academic research articles are more similar with one another than with the text type of public journals, which is different from them with different characteristics.

The factor scales plot of Factor 3 also suggests that the text types are distinguished regarding this factor, but in a different way. The text types are arranged so that the text type of introductory textbooks has the highest position on this plot with some distance to the other two text types, suggesting that in consideration of Factor 3 the texts of introductory textbooks have different characteristics than the texts of academic research articles and public journals. The reason for this difference will be analysed in the next section where the communicative functions shared by the co-occurring linguistic features on each factor are assessed.

On Factor 4 all three text types have almost similar factor scores and they are grouped close together. This suggests that the text types are not distinguished on this factor, but that they are rather similar regarding this factor. The reason for their closeness and similarity will be analysed in the next section.

This section has presented the results of the MD analysis to the first research question *Can the hypothesised text types be distinguished*? The graphic plots in Figure 4 show the text types in relation to one another with respect to Factors 1, 2, 3 and 4. Based on the plots it is suggested that to some extent the three text types are distinguished from one another. Regarding Factors 1 and 2 the text types are distinguished so that the text type of popular journals differs from the text types of introductory textbooks and academic research articles. The factor scales plot of Factor 3 also suggests that the text types are distinguished, but in consideration of this factor the texts of introductory textbooks are suggested to have different characteristics than the texts of academic research articles and public journals. With respect to Factor 4 the text types are not distinguished. Having similar factor scores on Factor 4 they are grouped within a close distance from one another. The reasons for the distinctions on Factors 1, 2, and 3 as well as for the similarity of text types on Factor 4 will be analysed in the next section.

As a summary to the first research question, the hypothesised text types can be said to be more or less distinguished with respect to the extracted factors. To answer the second research question also the communicative functions shared by the co-occurring features need to be assessed, so we continue to the next section to find out if software engineering texts written for different audiences show variation.

3.3.3 Linguistic dimensions vs. variation

The MD method is based on the assumption that frequently in texts co-occurring linguistic features have one or several communicative functions that co-occur, because they are linked to variation in the communicative situation (Biber 1988: 63-64). Four factors of co-occurring features were extracted in this study, and the text types and their relations were analysed in respect to these factors in section 3.3.2. The text types were suggested to be more or less distinguished from one another. Now we continue to assess the communicative functions of the features that constitute the factors and interpret them as linguistic dimensions. The dimensions are given names, labels, to indicate why the features co-occur in the text types.

The linguistic dimensions and their labelling offers a way to discuss text types and text type variation. In this section we answer the second research question *Do software engineering texts written for different audiences show variation?,* to find out if there is variation between the text types of software engineering written to three groups of audience.

The discussion here is based, firstly, on the plots of the factor scales structure above in Figure 4, and secondly, on the combinations and characteristics of the linguistic features that load on each factor. The complementary relationship between the positive and negative loadings is taken into consideration as well as the strength of the loads. The communicative functions shared by the co-occurring features in each factor are assessed, bearing in mind Biber's (1988: 92) comment that even though the co-occurrence patterns are extracted in a quantitative manner by factor analysis, the interpretation stage is tentative.

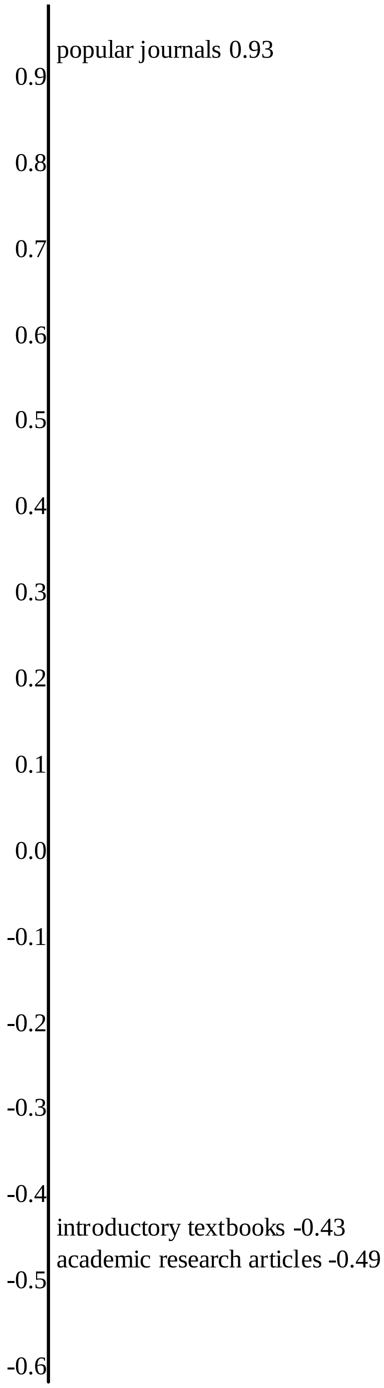
Because this study attempts, as far as possible, to adopt the MD method (Biber 1988), also the communicative functions of the linguistic features in this study are primarily taken from his study and its Appendix II (Biber 1988: 211-245).

Factor 1

The linguistic features on Factor 1 and its graphic plot are repeated for convenience.

Linguistic features on Factor 1

CONTRA	0.87	- contractions
2NDPER	0.85	- second person pronouns
3RDPER	0.49	- third person pronouns
ANANEG	0.48	- analytic negation
(TYPETO	0.39)	- type/token ratio
- - - - - - - - - - - - - - -		
PASSIV	- 0.67	- agentless passives
PREPOS	- 0.69	- total prepositional phrases
CONJUN	- 0.65	- conjuncts
WHRESU	- 0.61	- WH relative clauses on subject position
(WORDLE	- 0.31)	- mean word length



Factor 1

The highest factor score of the public journals text type on Factor 1, the first linguistic dimension, reflects high frequencies of contractions, second and third person pronouns, analytic negation, and a rather high type/token ratio, that is, the features with positive weights on Factor 1. It also reflects popular journals texts as having fewer occurrences of agentless passives, total prepositional phrases, conjuncts, and WH relative clauses on subject position, whereas the text types of introductory textbooks and academic research articles have the opposite characteristics.

The frequencies in Table 6 in section 3.3.1 also support this distribution. On the one hand, the features that have positive loads on this factor and that are suggested to characterise the text type of popular journals, likewise occur with the highest frequencies in the text type of popular journals. For instance, the mean frequency of contractions in the text type of popular journals is 9.9, whereas in introductory textbooks it is 0 and of academic research articles 0.1. There was not even one contraction in the text type of introductory textbooks and only one in academic research articles. Second and third person pronouns also occurred most frequently in popular journals texts, with mean values of 20.9 and 8.4, whereas there were fewer occurrences of them in introductory textbooks and academic research articles. Regarding analytic negatives the difference is not so significant, but, however, popular journals texts included more analytic negations that the other two text types.

On the other hand, all the features with negative loads on Factor 1, that is, agentless passives, total prepositional phrases, conjuncts, and WH relative clauses on subject position, occur in a complementary manner, that is, less frequently in the texts of public journals. In turn, they occur more frequently in the text types of introductory textbooks and academic research articles.

To give a label for this linguistic dimension we need to have a closer look at the communicative functions of the linguistic features on Factor 1. Contractions that have the highest weight are generally regarded as a feature of colloquial discourse. They are not preferred in formal, edited writing and they indicate a reduced surface form that signifies a generalised or uncertain presentation of information (Biber 1988: 243, 106). Analytic negation on Factor 1 is suggested to be more colloquial and fragmented than synthetic negation (Tottie 1983, as quoted by Biber 1988: 245). Analytic negation is also associated with fragmented presentation of information which may result in low content of information (Biber 1988: 106).

Features that are related to interpersonal communication, that is, second and third person pronouns, have high weights on this factor. Second person pronouns are often used in interactive discourse and third person pronouns have usually human referents outside the communicative situation at hand (Biber 1988: 105, 92).

The features with positive weights on Factor 1 can be suggested to have interactive and affective characteristics, marking interpersonal interaction, personal involvement and even colloquial discourse. A label of ´Interactive & affective focus´ is suggested to characterise the text type that has a high score on this factor, popular journals, based on the features with positive loads on Factor 1.

The features that have negative loads on Factor 1 reflect discourse which is more information oriented and formal. Passives are generally associated with formal discourse and style. Discourse with many passive constructions is typically abstract and technical, and when the agent is dropped altogether, as in the case of this study where the agentless passives were counted, it results in a static, abstract presentation of information (Biber 1988: 112). Conjuncts and prepositions are regarded to be associated with high information content (Biber 1988: 237, 239).

The features with negative weights on Factor 1 are suggested to have formal, informational characteristics, with abstract and even technical content. Thus a label ´Informational & abstract focus´ is suggested for the texts with a low score on Factor 1, introductory textbooks and academic research articles, based on their characteristics of features with negative loads on Factor 1.

We can have a look at text samples to illustrate the linguistic features on Factor 1. In sample 14, extract from text PJVAR04, the features with interactive and affective focus, that is, contractions, second and third person pronouns (no analytic negation is available in the same text), are marked in bold.

14) **HELP´S** ON THE WAY
Personal finance management software can liberate people like these from much of **their** guilt and **their** worry. It can also save **them** time and often money. Better yet, **there´s** little learning involved: Our nine review packages - Balance Point 1.1a, CheckFree 3.0, "Cheque-It-out" 2.0 (a Windows version of "Cheque-It-Out" was not ready in time for our review, but it is shipping now), Managing **Your** Money 9.0, Microsoft Money 2.0, MoneyCounts 7.0, Quicken for DOS 6.0, Quicken for Windows 2.0, and WinCheck 3.0p - all use a simple checkbook metaphor where information entered in an onscreen check is automatically transferred to the check register (or vice versa). When all transactions are entered, the program prints the checks and the information becomes part of a permanent, detailed record-keeping scheme that lets **you** see exactly how **you´re** spending **your** money by way of charts of diagrams. Some packages go further, providing links to electronic bill-paying services that pretty much eliminate check writing altogether. (Text PJVAR04)

The sample includes three contractions, four second person pronouns and three third person pronouns. The text gives an impression of interaction between the writer and the reader. The style is informal, possibly caused by the many contractions and speech like expressions such as *pretty much!*, *you see exactly how you´re spending your money …*

In sample 15 below of the text type of introductory textbooks we can find several features with negative loads on Factor 1, that is, agentless passives, total prepositional phrases, conjuncts, and WH relative clauses on subject position. These characteristics are suggested to have informational and abstract focus.

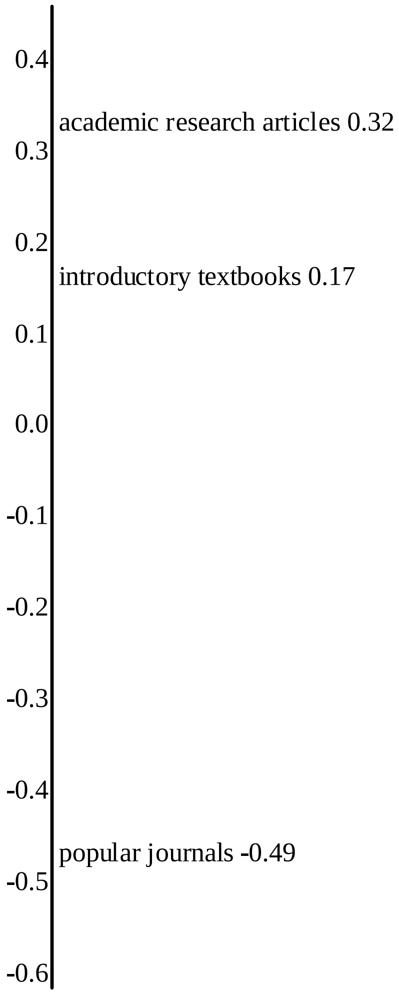
15) One important feature **of** top-down design is that **at** each level the details **of** the design **at** lower levels are **hidden**. Only the necessary data and control **which** must be **passed** back and forth over the interface are **defined**. **Furthermore**, if a data structure is **contained** wholly **within** a lower-level module, it need not be **specified** until that level is **reached in** the design process. **However**, if data must be **shared** by several modules **at** some level, then the data structure must be **chosen** before progressing **to** a lower level. The design will include both the data structure and the means **of** data access **for** each involved module. (Text ITVAR13)

Several passives, conjuncts and prepositions make it sound formal and official, with a high information content. Sample 16 below gives also the same feeling of official discourse, with the added research article likeness.

> 16) Some initial research has been **conducted in** the area **of** applying intelligent support **to** GDSSs including an automated group conflict resolution system [64], a system which actively filters and structures information exchange [23], an intelligent agent which monitors session activity [16,20], and intelligent electronic mail management systems to filter messages [9,37,48]. **In addition**, many attempts have been **made** over the years to implement automated natural language processing. **However**, most **of** these implementations have been **restricted to** a particular domain (such as the understanding **of** legal terminology or toy block movements) [7, 8, 21, 27, 38, 39, 42, 61, 62, 68, 72]. (Text ARVAR18)

Based on the communicative functions of the linguistic features on Factor 1, a label of ‘Interactive and Affective Focus vs. Informational, Abstract Focus’ was given for Dimension 1. Dimension 1 is suggested to characterise the differences and show variation between the text types of popular journals as compared to text types of introductory textbooks and academic research articles. Popular journals texts are suggested to be to some extent fragmented, affected and even colloquial, whereas the text types of introductory text books and academic research articles are regarded as more informational, abstract and even technical. The distribution of the linguistic features on Factor 1 and its graphic plot give support to the labelling and the interpretation that Dimension 1 shows variation between the text types of this study according to their interactive and affective characteristics versus informational and abstract, even technical content.

Factor 2



Factor 2

Linguistic features on Factor 2

THAVCO	0,89	- *that* clauses as verb complements
THADEL	0,72	- subordinator *that* deletion
PRIVER	0,45	- private verbs
- - - - - - - - - - - - -		
NOUNS	- 0,53	- total other nouns

There are three features with a positive and one with a negative load on Factor 2. Factor 2 repeats the same relationship as on Factor 1 that the text type of popular journals is separated from academic research articles and introductory textbooks, which are closer together. It is suggested that with respect to Factor 2 they are more similar with one another than with the public journals texts. However, the assessment of the communicative functions and labelling of this factor is not as straightforward as for Dimension 1.

The only feature with a negative load on Factor 2 is total other nouns. Based on Factor 2 and its graphic plot the text type of public journals is suggested to include high frequencies of various nouns, while, in a complementary pattern, the text types of academic research articles and introductory textbooks should reflect infrequent occurrences of total other nouns. However, the statistics in Table 6 do not directly support this, because the frequencies of nouns in the three text types are very similar as reported in section 3.3.1. The graphic plot indicates that the text type of public journals has the highest frequency of nouns, but, in fact, according to Table 6 the text type of academic research articles has the highest frequency with a mean value of 223 per text, public journals the second highest mean of 218, and in introductory textbooks the mean is 212. A high frequency of nouns in discourse is suggested to indicate a lot of abstract information, instead of interpersonal or narrative concern (Biber 1988: 227). Because all three text types have almost as high occurrences of total other nouns, they can all be regarded as highly informational. We have come to this conclusion earlier, because the topic of all text types is the science of software engineering, as discipline, subject and topic. The informational content cannot thus be a decisive, discriminating characteristic, and no distinction or variation between the text types can be based on the amount of informational content only.

Academic research articles and introductory textbooks, the text types with the highest scores on Factor 2, are suggested to be characterized by the features with the positive loads on Factor 2, that is, frequent occurrences of *that* clauses as verb complements, subordinator *that* deletion, and private verbs. The text type of popular journals which has a lower score on the factor scales plot of Factor 2, is claimed to have the opposite characteristics, that is, fewer occurrences of features with the positive loads.

In section 3.3.1 low frequencies of subordinator *that* deletion and *that* clauses as verb complements were observed in Table 6 to be small and similar in all three text types, and the mean requencies of *that* clauses as verb complements were likewise small. In addition, it was observed that private verbs occur with almost similar frequencies in all text types.

The frequencies are very small, so even though the graphic plot of the text types in relation to each other basically confirms the positive loads of features subordinator *that* deletion and *that* clauses as verb complements and their mean frequencies based on Table 6, except for private verbs and total other nouns, it is not reasonable to suggest any label for this factor. More samples and data are needed to verify this dimension, and the communicative characteristics and functions of the features also need further study.

We can, however, have a look at sample 17 from the text type of academic research articles (Text ARVAR24). It shows the features with the positive loads on Factor 2, *that* clauses as verb complements, subordinator *that* deletion, and private verbs, marked in bold. The feature with a negative load, total other nouns, is also marked.

> 17) When reading jointly, each **individual** may skim, **assuming that** others are reading more thoroughly, or will pick up what they may have missed.
> Finally, we **found** no **effects** of EMS on **member** assessments of their **meetings**. Although this **result** is consistent with our **hypothesis**, we **think** it should be interpreted cautiously because the reliability of some of these **measures** was low. The satisfaction **measures**, adapted from Schein´s (1969) **work**, had high reliability, and it **appears** that EMS had no **effect** on these **measures**. The **measures** derived from Hackman (1989) similarly were not affected by EMS, but the reliability of those **measures** was low. It **appears** that low **structure** EMS has little **effect** on **user** satisfaction with **task** and interpersonal **processes**, but this needs further investigation, with more reliable **measures**. (Text ARVAR24)

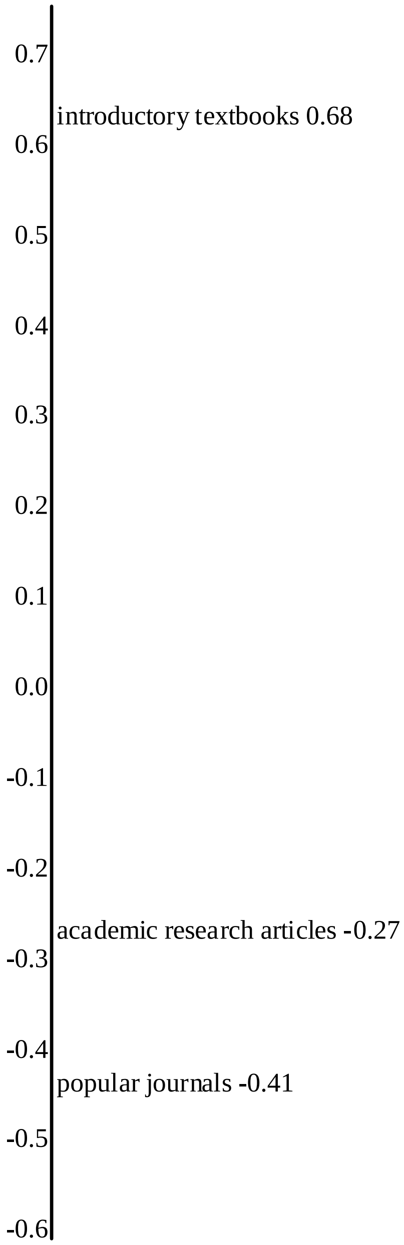
The text has many nouns, which supports the claim that a high frequency of nouns bring of a lot of abstract information into discourse.

The few linguistic features on Factor 2 had so small and similar frequencies that they did not give support for variation regarding the text types on Factor 2, and no attempt was made to give a label for it. In addition, the distribution of the total other nouns feature was contradictory to the usual factor interpretation, possibly because all three text types can be regarded as highly informational which may distort the findings. Further research is needed to confirm Factor 2 and give it further clarification.

Factor 3

Linguistic features on Factor 3

PRESTE	0,69	- present tense
PLACAD	0,48	- place adverbials
DOPROV	0,45	- pro-verb *do*
(CONJUN	0,36	- conjuncts)
- - - - - - - - - - - - -		
PASTTE	- 0,70	- past tense
(PREPOS	- 0,49	- total prepositional ph
(PRIVER	- 0,39	- private verbs)



Factor 3

Texts with high scores on this factor, that is, the text type of introductory textbooks, is suggested to be characterised by the features with positive loads, that is, present tense, place adverbials, pro-verb *do*, and conjuncts, with fewer occurrences of past tense, total prepositional phrases and private verbs. The feature that has the highest negative load on Factor 3 is past tense. The factor scales plot suggests that regarding Factor 3 the text type of introductory textbooks is different from the text types of academic research articles and popular journals, which in turn are more similar with one another.

Table 6 with the frequency statistics supports the above suggestion. The texts of introductory textbooks have the highest frequencies of features with positive loads on this factor, that is, of present tense, pro-verb *do,* and conjuncts. Place adverbials occur in public journals texts with the same small frequency. The texts of introductory textbooks have the lowest frequency of past tense, which gives further support to Factor 3 and its plot. Contrary to Factors 1 and 2, the text type of introductory textbooks is separated from the text types of academic research articles and public journals regarding Factor 3. One reason for this may be the distribution of the present tense feature. Academic research articles have the lowest frequency of present tense, whereas, as mentioned, introductory textbooks have the highest. Regarding the present tense the text types are suggested to be different, but as the sample size is small, it may lead to a statistical distortion, where the loading of few features may come out disproportionately.

Factor 3 includes several verbal features such as present and past tense, pro-verb *do,* and private verbs, thus indicating verbal, instead of nominal, style. Present tense which has the highest positive load refers to actions occurring in the immediate context of interaction (Biber 1988: 105). Another feature that refers to locations nearby is place adverbials. They are used for place referents in the actual physical context of the discourse (Biber 1988: 110). Pro-verb *do* is suggested to signify smaller informational focus, due to processing constraints or a higher concern with interpersonal matters (Biber 1988: 226). This also conveys closeness. These positive features of Dimension 3 are labelled as ´Immediacy of context´.

The feature with the highest negative load on Factor 3 is the past tense. Past tense forms are regarded as markers of narrative texts (Biber 1988: 223). Regarding the frequencies of past tense forms in Table 6, they are clearly the highest throughout the texts of public journals and there are three texts of academic research articles which have high frequencies of past tense. This may be another reason which explains why the texts of public journals and academic research articles are grouped so close together on the factor scales plot of Factor 3, indicating their similarity, even though regarding Dimension 1 they were different.

The other two features with negative loads on this factor are total prepositional phrases and private verbs. As mentioned earlier, prepositions are considered to bring information into discourse, whereas private verbs express, for instance, private attitudes, thoughts and emotions. These features coincide with characteristics that are understandable as typical of journal texts, narrative style with informational and possibly also personal and interactive concerns.

Thus a label of ‘Immediacy of context vs. Narrative concerns’ is suggested for Dimension 3.

To illustrate the linguistic features of Dimension 3 sample 18, from the text type of introductory textbooks, shows the positive features on Dimension 3 marked in bold, that is, present tense, place adverbials, pro-verb *do*, and conjuncts.

> 18) The functional specification **states** what the problem **is**. Unless a declarative approach to programming **is** to be used, or an application package **is** available (for example **see** *application generators*, **below**), we now **have** to solve the problem by finding a solution; that is, derive the how. A question that should be asked **is** 'has the problem in whole or part been solved before?'. If it has, 'could we use or amend a previous solution?'. In this context an application package might be attractive. If a package **is** feasible this can result in savings both in cost, time and the recurring cost of maintenance of one's own specific software solution. Use of a package can reduce or eliminate the design and programming steps in the project's life-cycle. Packages **are** discussed **below** in Sec. 7.3.
>
> 7.1.5 Design
>
> The functional specification **defines** the processes and their inputs and outputs, that is, what the system **is** to **do**. (Text ITVAR10)

The above text is written in the present tense, and it gives a feeling of immediacy, closeness between the writer and the audience. It includes clear instructions of how to find solutions by asking questions, etc., as can be expected of a textbook text.

Sample 19 of the text type of public journals is marked with the features with negative loads on Dimension 3, that is, past tense, total prepositional phrases, and private verbs.

> 19) Performance differences **among** these products **were** subtle, **in** part because we **were** passing a fairly small (10K) file. But if we **increased** the file size, (say, **to** several megabytes), **used** a lower speed link (say, 56 kilobits **per** second), or **separated** the two hosts **with** several routers, these minor performance differences would become extremely important (**for** more information, **see** "Performance Tests: TCP/IP Packages **for** NetWare 3.11").
> We **ran** three tests **on** each product. Two tests **measured** the time it **took** to transfer a file **from** the DOS client **to** the Sun host, and vice versa, using FTP. Our third performance test **looked at** the emulation support **for** the CAT command, a standard Unix command used to view a document **on** a remote station **via** a Telnet session, which is similar to the TYPE command **in** DOS. (text PJVAR03)

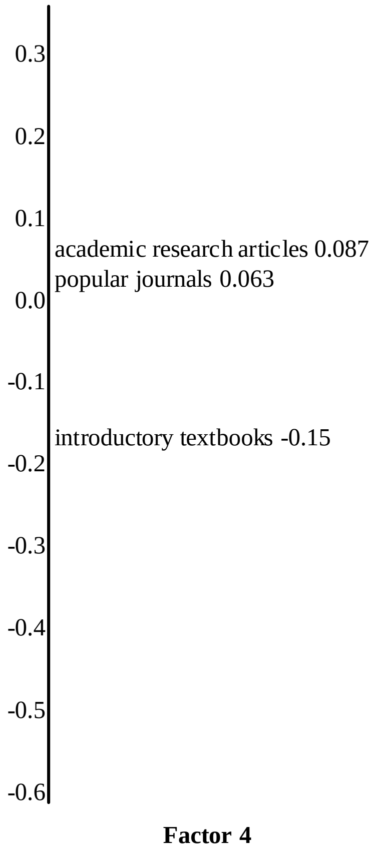
The sample reminds of narration, perhaps because of the past tense form, the writer using *we* as the narrator, the use of short clauses and *say*, with a lot of punctuation resembling pauses that a speaker would have while thinking about what to say. This sample shows again how informative the texts are.

A label of ´Immediate context vs. Narrative concerns´ was suggested for Dimension 3. It seems to support variation between the text type of introductory textbooks versus public journals and academic research articles, but further research is required to confirm it.

Factor 4

Linguistic features on Factor 4

WORDLE	0,73	- mean word length
TYPETO	0,70	- type/token ratio
DEMONS	0,68	- demonstratives
DEMPRO	0,55	- demonstrative pronouns
(WHRESU	0,35)	- WH relative clauses on subject position
PERFAS	0,31	- perfect aspect
- - - - - - - - - - - - -		
(NOUNS	- 0,41)	(- total other nouns)



The text types have similar factor scores on the plot within a close distance from one another, suggesting that they are similar regarding the features on Factor 4.

Similarity is supported by Table 6 in section 3.3.1, where the mean word length was found to be either 5 or 6 ortographic letters in all text types. The mean value of the type/token ratio, the number of different words in a text as per centage, is also almost similar in all text types (AR 49.0, IT 44.3 and PJ 54). Both long words and high type/token ratio are regarded to have informational focus.

The other features with positive loads have also similar frequencies in all text types. Demonstratives have the following mean frequencies per texts of 1,000 words; AR 10.4, IT 8.9, and PJ 9.4, and demonstrative pronouns; AR 3.4, IT 4.6 and PJ 4.2, and WH relative clauses on subject position; AR 1.8, 3.1, and PJ 0.2.

Based on the features on Dimension 4 we can suggest a label for it. The positive loads of mean word length and type/token ratio signify informational content, as well as the negative load of total other nouns. As for perfect aspect verbs, as mentioned earlier, past tense forms are regarded to signify narrative concerns. Demonstratives as well as demonstrative pronouns are associated with informal, unplanned types of discourse (Biber 1988: 113). So the linguistic features of Dimension 4 have communicative functions related to informational, even informal and unplanned discourse, with narrative features. So label ´Informational focus with narrative concern´ is given for Dimension 4. Dimension 4 cannot be suggested to show any variation between the text types, because of their similarity regarding the linguistic features that comprise it.

In this section the factor scores and the relations of the text types on factors were analysed by having a look at the combinations of features and their communicative functions. As an answer to the second research question, *Do software engineering texts written for different audiences show variation?*, variation was found regarding one dimension. Dimension 1, ‘Interactive and Affective Focus vs. Informational, Abstract Focus’, was suggested to show variation between the text types of popular journals as compared to text types of introductory textbooks and academic research articles.

The few linguistic features on Factor 2 did not give support for labelling or variation between the text types regarding this factor. Dimension 3, ´Immediate context vs. Narrative concerns´ was to some extent found to support variation between the text types, but further research is required to confirm it. Regarding Dimension 4, ‘Informational focus with narrative concern’, the text types were so similar that no variation was established between them.

4 DISCUSSION AND CONCLUSION

In this chapter I summarise the purpose and method of the study with the main results, compare the results with other studies, evaluate the study, and suggest further research.

The purpose of this study was to find out if there is variation in the scientific text types of software engineering written for different audiences.

The sample texts were divided in three text types; academic research articles, introductory textbooks and popular journals. The audience was divided likewise into three groups; academic, introductory, and laypersons. The study combined textual and variation analysis with quantitative measures. The method was provided by the MD approach and the concept of dimension of linguistic variation. The co-occurrence relations among the linguistic features of sample texts were grouped into factors, which were then interpreted as linguistic dimensions, based on the assessment of their communicative functions. The dimensions were given tentative labels to signify why the features co-occured in different text types.

Dimension 1 was given the label ‘Interactive and Affective Focus vs. Informational, Abstract Focus’. The distribution of the linguistic features and its graphic plot gave support to the labelling and the interpretation that Dimension 1 shows variation between the text types of this study according to their interactive and affective characteristics versus informational and abstract, even technical content. It was suggested to characterise the differences between the text type of popular journals as compared to text types of introductory textbooks and academic research articles. Popular journals texts were suggested to be to some extent fragmented, affected and even colloquial, whereas the text types of introductory text books and academic research articles were regarded as more informational, abstract and even technical.

The few linguistic features on Factor 2 had so small and similar frequencies that they did not give support for variation regarding the text types on this factor, so no attempt was made to give a label for it. All text types of this study being informational may have distorted the findings. Variation between the text types regarding this factor was suggested to need further research.

The tentative label of ‘Immediate context vs. Narrative concerns’ was given for Dimension 3, which seemed to support the interpretation that there was some variation between the text type of introductory textbooks vs. public journals and academic research articles, but further study was suggested to confirm it.

The label ‘Informational focus with narrative concern’ was suggested for Dimension 4, but the text types were so similar regarding it that their characteristics were not sufficiently differentiated. So Dimension 4 was not considered to show any variation between the text types.

In summary we may conclude that variation found on Dimension 1 indicates that texts and discourse written for laypersons, that is, the text type of popular journals, approaches non-scientific discourse. I would like to suggest that the differences in audiences may be regarded as different communicative situations, which result in differences in the discourse.

This study was based on the MD method (Biber 1988). Because the corpus and the number of linguistic features was smaller in this study than that of Biber, we cannot make any direct comparisons between them. However, we can compare the mean frequencies of the linguistic features per texts of 1,000 words. See Table 7 below, which shows the mean frequencies of the linguistic features in this study and that of Biber (1988: 77-78).

Table 7. Frequencies of linguistic features in this study and that of Biber (1988)

Linguistic	Mean values in	
feature	1) this study	2) Biber (1988)
1. PASTTE	8.5	40.1
2. PERFAS	3.1	8.6
3. PRESTE	60	77.7
4. PLACAD	0.7	3.1
5. TMEAD	2.2	5.2
6. 2NDPER	7.7	9.9
7. 3RDPER	5.9	29.9
8. DEMPRO	4.2	4.6
9. DOPROV	0.5	3
10. NOUNS	214.3	180.5
11. PASSIV	16.2	9.6
12. THAVCO	2.7	3.3
13. INFINI	19.1	14.9
14. WHRESU	1.7	2.1
15. WHREOB	0	1.4
16. PIEDPI	0.8	0.7
17. PREPOS	106.2	110.5
18. TYPETO	48.7	51.1
19. WORDLE	5.1	4.5
20. CONJUN	6.7	1.2
21. DEMONS	9.7	9.9
22. PREDMO	5.6	5.6
23. PRIVER	9.1	18
24. SUAVER	3	2.9
25. CONTRA	3.5	13.5
26. THADEL	0.2	3.1
27. ANANEG	5.4	8.5

Columns 1) and 2) in Table 7 show that in the two studies the mean frequencies of linguistic features per texts of 1,000 words are to a large extent very similar. There is one mean frequency that is exactly the same in both studies: prediction modals (No. 22 PREDMO) with a mean value of 5.6. Some mean values of frequencies are close to each other. For instance, the mean frequency of demonstrative pronouns (No. 8 DEMPRO) is 4.2 in this study, compared with that of 4.6 in Biber. The mean frequency of total prepositional phrases (No. 17 PREPOS) is 106.2 in this study, whereas in Biber's study it is 110.5. The mean frequency of demonstratives (No. 21 DEMONS) is 9.7 in this study, compared with that of 9.9 in Biber. In this study the mean frequency of type/token ratio (No.

18 TYPETO) is 48.7 compared with that of 51.1 in Biber, and the mean word length (No. 19 WORDLE) in this study is 5.1 compared with that of 4.5 in Biber.

The biggest differences in the mean frequencies of the linguistic features of this study and that of Biber’s (1988) are in nouns and the past tense forms. In the present study nouns occur frequently, with a mean of 214, whereas in Biber’s study there are 180 occurrences per 1,000 words. We may consider this to support our earlier observation that all the text types of this study are highly informational. Past tense forms occur in this study with a mean frequency of 8, whereas in Biber’s study the corresponding value is 40.

In summary, on the basis of Table 7, we can say that the mean values of frequencies coincide to large extent. There are differences, which is quite natural based on the differences in corpus, but as a small test on the credibility of the method used this comparison supports the analysis so that the mean values can be considered to indicate some distribution of these features in English texts.

Now we proceed to relate the results of this study to other studies performed by using the MD method. In his study of expository prose Grabe (1987) found support for his labelling of four textual dimensions which discriminate text types and his results suggested that expository prose is a genre with identifiable sub-types. Atkinson (1992, 1996), in turn, was able to claim that medical research writing has changed gradually over the years, in contrast to the more general view of it having changed in more sudden steps resulting from changes in the scientific methods. Taking into consideration the limitations of this study, support was found likewise for the finding of one dimension to show variation between text types

Myers (1989) suggested that even though scientific writing does not involve direct personal contact between the writer and the audience, it is interaction that makes the scientific writer use complicated forms to hedge claims or to criticise. The significance of the interaction between the writer and reader of scientific writing is worth and requires further research. Based on Widdowson’s view on scientific journalism approaching non-scientific discourse and the findings of this study, I would suggest that public journals texts are more oriented towards interaction than information, which makes them different and causes variation in relation to other text types.

The concept of communicative function was not specified in this study. It turned out to be of significant importance for the interpretation of the findings, and additional background information would have been valuable. There are contrasting views about the term communicative function. For instance, Beedham and Bloor (1989) claim that the formal realisation of the communicative function is not grammatical, but lexical items function as markers of the communicative function. Thus further information and research on this issue is required.

The present study used a quantitative method for variation analysis, which is not that common in the field of ESP. The size of the corpus and the number of variables had to be limited because no tagging program was available to simplify and shorten the coding of the linguistic variables. In such a manual coding process, a second coder would be useful also increase coding reliability. The size of the corpus limited the possibility to have better justifiable and more comparable findings with other studies.

With a small corpus a few particular variables may distort the factor structure. In this study an additional research method, for instance, a rhetorical analysis would have provided additional information for the interpretation of the findings. Sample groups of additional text types of different special fields would have helped to overcome the difficulty caused by all the text types being informative. All texts had high frequencies of features associated with informative content which complicated their discrimination. It would also have been beneficial for the study to include more background information about communicative functions and their interpretations, as well as about variation studies.

Despite the limitations, the theoretical framework provided a good basis for this study and the research questions were given tentative answers. The purpose of this study was to examine scientific English and text type variation by the MD method. Characterisations and descriptions of the text types of this study could be given.

The MD method can be recommended for further research on various text types and their linguistic features, to generate linguistic descriptions for the requirements of ESP or EST. The method is an escpecially good tool in corpus linguistics with large amounts of computerised data. By using large amounts of samples of natural language you can examine linguistic features and characteristics that are really used by the speakers and writers of various special fields. The linguistic descriptions can be used, for instance, for the creation of appropriate language instruction methods and materials for different target groups, the development of computer aided translation techniques, and the generation of online dictionaries.

On the basis of the findings of this study I would like to suggest a research where the focus is more specifically on the interaction between the writer and reader of scientific writing, to find out in more detail how the orientation of texts towards interaction makes them different and causes variation in relation to other text types.

BIBLIOGRAPHY

Atkinson, D. 1992. The evolution of medical research writing from 1735 to 1985: the Case of the Edinburgh Medical Journal. *Applied Linguistics* 13, 337-374.

Atkinson, D. 1996. The philosophical transactions of the Royal Society of London, 1675-1975: A sociohistorical discourse analysis. *Language in Society* 25, 333-371.

Barber, C. L. 1962. Some measurable characteristics of modern scientific prose. In Swales, *Episodes in ESP*. Hemel Hempstead: Prentice Hall, 1-14.

Beedham, C. and M. Bloor 1989. English for computer science and the formal realization of communicative functions. *Special Language Fachsprache* 1-2, 13-24.

Biber, D. 1986. Spoken and written textual dimensions in English: resolving the contradictory findings. *Language* 62 (2), 384-414.

Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Grabe, W. 1987. Contrastive rhetoric and text-type research. In U. Connor, and R. B. Kaplan (eds.), *Writing across languages: Analysis of L2 text*. Reading, MA: Addison-Wesley, 115-137.

Halliday, M. A. K. 1988. On the language of physical science. In M. Ghadessy (ed.), *Registers of written English: Situational factors and linguistic features*. London: Pinter, 162-178.

Halliday, M. A. K. and R. Hasan 1985. *Language, context and text: aspects of language in a social-semiotic perspective.* Victoria: Deaking University.

Hutchinson, T. and A. Waters 1981. Performance and competence in English for Specific Purposes. *Applied Linguistics* II, 56-69.

Hutchinson, T. and A. Waters 1987. *English for specific purposes: a learning-centred approach.* Cambridge: Cambridge University Press.

Johns, A. M. and T. Dudley-Evans 1991. English for specific purposes: international in scope, specific in purpose. *TESOL Quarterly* 25, 297-314.

Lackstrom, J., L. Selinker and L. Trimble 1973. Technical rhetorical principles and grammatical choice. *TESOL Quarterly* 7, 127-136.

Myers, G. 1989. The pragmatics of politeness in scientific articles. *Applied Linguistics* 10, 1-35.

Quirk R., S. Greenbaum, G. Leech and J. Svartvik 1985. *A comprehensive grammar of the English language*. New York: Longman.

Strevens, P. 1977. Special purpose language learning: A perspective. *Language Teaching & Linguistics Abstracts* 10 (3), Cambridge: Cambridge University Press, 145-163.

Strother, J. B. and J. M. Ulijin 1987. Does syntactic rewriting affect English for science and technology (EST) text comprehension. In J. Devine, P. L. Carrell and D. E. Eskey (eds.), *Research in reading in English as a second language*. Washington, DC: TESOL, 91-101.

Swales, J. 1981. The function of one type of particle in a chemistry textbook. In L. Selinker, E. Tarone and V. Hanzeli (eds.), *English for academic and technical purposes. Studies in honor of Louis Trimble.* Rowley, MA: Newbury House, 40-52.

Swales, J. (ed.) 1988. *Episodes in ESP*. Hemel Hempstead: Prentice Hall.

Swales, J. M. 1990. *Genre analysis: English in academic and research settings.* Cambridge: Cambridge University Press.

Tarone, E., S. Dwyer, S. Gillette and V. Icke 1981. On the use of the passive in two astrophysics journal papers. In Swales, *Episodes in ESP.* Hemel Hempstead: Prentice Hall, 188-205.

Widdowson, H. G. 1979. *Explorations in Applied Linguistics*. Oxford: Oxford University Press.