Classification along Genre Dimensions

Exploring a Multidisciplinary Problem

Mikael Gunnarsson

2011

UNIVERSITY OF BORÅS
SWEDISH SCHOOL OF LIBRARY AND INFORMATION SCIENCE

UNIVERSITY OF GOTHENBURG
GRADUATE SCHOOL OF LANGUAGE TECHNOLOGY

Dissertation at
Swedish School of Library and Information Science
University of Borås

Mikael Gunnarsson
Classification along genre dimensions

ISBN 978-91-89416-27-9
Cover: Jennifer Tydén
Printed: Intellecta Docusys, Göteborg, 2011
Series: Skrifter från Valfrid, nr. 46
ISSN 1103-6990
Typeset by the author using LaTeX
Available in PDF version at `http://hdl.handle.net/2320/7920`

Abstract

This thesis treats the sociotechnical notion of genre as a conflation of a communicative situation and a community of practices involved in producing and using documents. It explores the ways in which documents may be mapped to the sociocultural contexts from which they emanate. In other words, it is concerned with the classification of documents along genre dimensions, with the purpose of supporting information seeking.

The thesis positions itself within Library and Information Science in two parts. Firstly, a theoretical framework for classification along genre dimensions is developed based on relevant theories and practices from Library and Information Science, as well as from sociologically motivated Linguistics, and neighbouring domains. Secondly, a setup for experiments, including feature derivation and reannotation of existing corpora, is designed in order to explore the relationship between text documents and genres, and the extent to which a mapping of documents to genres can be realized in real world applications.

The experimental part of the thesis relies on an existing corpus for genre classification research, used in comparable research, with an addition of a slight extension. In the experiments, combinations of feature sets and target genres are evaluated, using traditional performance estimators for classification performance.

The outcome of the first part of the work indicates that the notion of genre with respect to classification is largely undertheorized in Library and Information Science. We need to know more about the nature of different genres, how to robustly identify the documents of a genre, and the impact genres have on information seeking. Inter-

disciplinary collaborative research would be most beneficial in these efforts. The results of the experiments of the second part are fairly inconclusive for the evaluation of feature sets, but it can be concluded that the optimal combination of feature sets and target genres is a crucial issue for high performance, and worthy of more investigation.

Sammanfattning

Utgångspunkten för den här avhandlingen har varit att en genre motiveras av en kombination av en kommunikativ situation och en social gemenskap i vilken dokument spelar en viktig roll. Avhandlingen utforskar begreppet genre med avseende på hur det används i samband med klassifikation, och då särskilt med hänsyn till tillämpningar för s k automatisk klassifikation.

Avhandlingens första del påvisar att de lingvistiska begreppen *register*, *texttyp* och *talakt* hänger samman med begreppet genre i egenskap av att vara språkliga typifieringar, och att de innehållsmodeller som utvecklas för skilda XML-tillämpningar kan återföras på genrebegreppet. Det framhålls att förståelsen av genrer såsom uttryck för social handling inte ges en särskilt framskjuten betydelse i kontexten av klassifikation på bibliotek eller i forskningsprojekt med uttryckligt fokus på automatisk genreidentifikation. En förklaring till detta är att klassifikation av naturen måste utgå ifrån observerbara och extraherbara särdrag i ett dokument. Det är därmed viktigt att hålla isär klasser av dokument som kan återföras på en genre och genrerna i sig, och att vara observant på att de namn som ges till klasser av dokument inte alltid ograverat kan tas som namn på genrer.

I en andra del av den här avhandlingen har en experimentell miljö utformats för att undersöka hur tillförlitlig automatisk klassifikation kan förväntas vara med olika extraherbara särdrag och genreuppsättningar. Tre olika klassifikationsmodeller har i varierande utsträckning utnyttjats för detta ändamål: Support Vector Machines (SVM), k-nearest neighbor (k-*NN*) och *K*-means klustring. Dessa algoritmer har tillämpats på en existerande corpus som använts tidigare i

utvärderingar av automatisk genreidentifikation, *KI-04*. *KI-04* har i en del av experimenten utvidgats med ytterligare data för att möjliggöra en fördjupad undersökning. Vidare har, för såväl *KI-04* som för utvidgningen, tidigare icke prövade särdrag extraherats och utvärderats: verbklasser som ger uttryck för olika talakter samt särdrag relaterade till den interna dokumentstrukturen. Särskilt intresse har ägnats åt studier av hur dokument kan återföras på genrer som emanerat från vad som kan betecknas som mer eller mindre vetenskapliga gemenskaper, t ex artiklar i vetenskapliga tidskrifter, tekniska rapporter och didaktiskt ägnat material.

Det kan, utan större förvåning, konstateras att utifrån de experimentella data som varit tillgängliga så är antalet klasser till vilka en samling dokument skall mappas av stor betydelse. Ju fler klasserna är desto fler felklassningar gör en algoritm, men samtidigt är skillnaden mellan de genrer de antas emanera ur av stor betydelse. Att skilja dokument i vetenskapligt orienterade genrer från dokument i olika typer av diskussionfora från varandra, är i allmänhet tillfredsställande robust. Det kan också konstateras att det är värdefullt att kunna identifiera prototypiska dokumentexempel på förhand, för de genrer som är av intresse.

Det går att skönja en tendens till att ju fler särdrag som är aktiva i klassifikationsprocessen, desto bättre resultat kan också förväntas, men exakt vilka särdrag som är mest effektiva tycks alltid vara beroende av vilka genrer som är av intresse.

Vad beträffar de två grupper av relativt innovativa särdrag som studerats kan sägas att de för genrer inom vetenskapliga domäner inte kan fastställas ha någon avgörande betydelse.

Sammanfattningsvis, att fastställa till vilken genre ett dokument skall mappas är en relativt osäker uppgift, såväl för ett mänskligt intellekt som för en algoritm. Genreklassifikation, eller, mer korrekt, klassifikation utmed genredimensioner, är ett relativt nytt och outforskat intresseområde. Det saknas tillräcklig kunskap om t ex hur olika särdragsuppsättningar påverkar klassifikationens resultat med avseende på olika kombinationer av genrer definierade huvudsakligen med avseende på dokumentens sociokulturella roll. Vidare saknas också tillräcklig kunskap om vilka effekter en större medvetenhet om

och ett större utnyttjande av genreindelningar skulle ha i informationssökningssammanhang. Framtida forskningsansatser kan med fördel orienteras mot tvärvetenskapliga ansatser till att studera genredimensionell klassifikation.

Preface

Looking back. Starting a thesis is not difficult. Once you have been admitted into a Phd education and got your funding, despite the grinning faces of those who for some reasons do not wish you to get the opportunity instead of someone else, all you have to do is to start reading, thinking, experimenting and writing.

Looks fine, if you stick to the source time schedule, which, of course, seldom happens. Soon, you realize that there is another part of your life that calls for attention when you least want it to. Relatives die, other responsibilities get you to revise your source time schedule and, one day, your funded time suddenly is out. Suddenly, you need to find time to finish your thesis within small slots between your ordinary work duties.

I have been a teacher in Library and Information Science since 1992 when I was engaged by the chief responsible for 'Bibliotekshögskolan' as a teacher assistant with responsibilities for courses related to information technology and knowledge organisation.

Rather soon I became responsible as a teacher for courses on Internet technology and Internet resources. I introduced phenomena such as *gopher*, *telnet*, and *wais* (now largely forgot). I tried to show the students how these phenomena could be used as sources in information seeking tasks. The World Wide Web was in its infancy, but soon became the main form for data communication on the Internet, and I introduced it to the students who also were taught how to publish web pages. This was back in 1994.

From this point of view an interest in markup languages grew strong in me and I began the study on how such phenomena could be useful in other ways than just to design the visual appearances of web pages, which, I soon came to realize, is a fallacy. This is from where I ended up in a thesis on genres and classification, if anyone wonders.

Looking back on this there is much to regret, and what must be learned is that satisfaction does come from leaving a much too ambitious project behind, finished to the extent that one does not have to be ashamed for the result.

Acknowledgements

I am indebted to all who have been helpful in the course of this mission, by giving me input of different kinds, doing proofreading, supervising me etc. I do not dare enumerating everyone that have been important, being at risk of forgetting someone. I do believe that everyone who has been helpful in the course of my mission, is well aware of this fact and should know that I am greatly thankful for that they have been present. However, I need to mention all the supervisors who have been involved during the course of my mission: Lars Höglund, Barbara Gawronska, Sandor Dárányi, Tor Henriksen and Elena Macevičiūtė. I also wish to thank my fellow colleague and friend Mats Dahlström, who has, most of all, been a helpful friend, and finally, my wife Gunnel, and my two daughters, Ariella and Esmeralda.

This work has been mainly funded by the Graduate School of Language Technology at Gothenburg University.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis is about the classification of text documents. In library and information science (LIS) the word “document” is almost intuitively understood as denoting any object that carries text, images or any kind of data. The nature of documents and its defining characteristics has occupied many within LIS. But as is the case with the words knowledge and information, the meaning of the word ’document’ has been extensively discussed and remains a fairly debated issue among the more theoretically inclined writers within LIS (see, for instance, Briet, 1951, Buckland, 1991). Within LIS and library practices, classification is intimately related (some would say subordinate) to information seeking and supports the task of information seeking by *dividing large collections of documents into groups of similar documents*. The notion of document similarity (and dissimilarity) is central to classification and it is obvious that there are many different kinds of document property similarities that can form the basis for the grouping of documents. Authorship, the time of publication, and topical contents are only three examples.

More specifically, this thesis deals with the classification of text documents where the properties considered for grouping are supposed to reflect genre adherence, *where genre is understood as a conflation of a communicative situation, and a community of practices in which documents play important roles*. An article by Carolyn Miller (1994) marks a starting point for this social conception of genre. This tradition is sometimes referred to as the “new genre theory”, where genre is understood as “typified rhetorical actions based in recurrent situations” (Miller, 1994, orig. publ. 1984). This theory will be further elaborated in Chapter 2.

A set of documents is considered to adhere to the same genre if the roles these documents play in more or less similar communities are sufficiently similar. Most often, but not always, such groupings of documents are given names. Examples of names assigned to document classes that distinctively take part in genres are bibliographies, research reports, and encyclopediae. Thereby, names given to such classes of documents are often taken as labels of genres. This is, and has to be emphasized, a notion of genre that has little to do, if anything at all, with literary or artistic genres.

This introductory chapter will outline the problem area, specify the aims of the work behind this thesis, and explain its motivation. It ends with a short description of the thesis structure.

1.1 The problem and its domains

The classification of documents has been a core problem in library practices ever since libraries evolved as a kind of repositories of human memory, but especially so since the middle of the nineteenth century (cf. Miksa, 1992, p. 104). As LIS has grown out of library practices and their needs, classification also occupies a core position in LIS as an academic discipline.

Arranging physical items in a predictive order on shelves or other kinds of storage utilities may be one of the most well-known examples of a classification task in libraries. It is one of the primary tasks when the number of items in a collection increases beyond a certain threshold and other people than the organisers themselves are expected to find items in the collection. According to which principle this should be done is, however, not self-evident. If all items are books and each of these books has one author, books can be arranged in an alphabetical order according to the initial letters of the authors’ family names. This is, generally speaking, a classification by way of grouping together books which are similar by virtue of the names of their authors.

In library practices this classifying principle is useful to some extent, but far from satisfying when the range of possible types of information access problems is considered. If someone wants and expects to find e.g. a treatise on Roman history, it would then be necessary to know in advance which authors have written treatises on Roman history. Libraries therefore need tools and principles that support many different points of departure for information seeking.

From the nineteenth century and onwards library practices have adopted many different *classification schemes* designed for the explicit purpose of organising books and other documents in libraries, by means of providing a “controlled vocabulary” for the designation of the contents of documents.[1] These schemes require that a librarian performs an analysis of the contents of the documents to be classified and assigns codes or other designators to the documents — designators that are enumerated in the schemes. Thereby the grouping of documents on the shelves can be based on these assigned designators and the structure of the classification scheme. Classification can thus also be approached as a descriptive process. This twofold character of classification is further elaborated in Chapter 3.

Of course, for every educated library practitioner, both authorship grouping and the use of classification schemes present well known principles and have for a long time been satisfactory for the organisation of library material. However, technological changes have increased the diversity of the *kinds of documents* relevant for library practitioners and library patrons, as well as the amount of documents that need to be organised within a certain restricted time span. The demand on library practices to keep pace with the patrons’ demands increases the need to find ways of using technology to organise collections of documents in a more timely and efficient fashion. In fact, documents in digital form are not organised on shelves, but need to be stored and organised on computer storage media, and visualized on a computer screen or in other kinds of media. In addition, when docu-

[1] The most well-known domain-independent scheme, from an international perspective, is probably the Dewey Decimal Classification scheme, designed by Melvil Dewey and published in its first edition in 1876 (Feather & Sturges, 1997).

ments are digital, metadata such as author names can be identified and extracted by algorithmic[2] means, which greatly increases the amount of documents that can be organised within a certain time span.

This situation necessitates that library practices take account of issues common to computer scientists, that is, finding algorithms that allow for a computer program to do some of the work. This fact brings part of LIS in close connection with computer science. The problem adressed in this thesis, the classification of document collections, thus realizes an intersection of LIS and computer science, where such research areas as *information retrieval*, *text categorization*, and *machine learning* are to be found.

The task of authorship classification (i.e. a simple alphabetical arrangement) is a fairly trivial task even for computers, given that you can specify rules for the robust identification and retrieval of author names in a digital document. In library practices, however, classification by means of predefined classification schemes is a more complex task that involves the analysis of the text within a document (or parts of this text) in order to determine its contents in terms of e.g. its topics (or its subject matter) — i.e. what it is about. Such a classification task involves, firstly, the human interpretation of words, clauses and larger entities. It implies the assignment of meaning to the text — the understanding of the use of language and its application to a communicative context. Secondly, it involves the translation of an analysis into the terminology or symbol system of the classification scheme (or the indexing language[3]). Thereby, the structure of these classifica-

[2] In this work the word *algorithm* refers to well-defined procedures for certain tasks that always arrive at a solution. Such a solution is always correct with reference to the algorithm, but not necessarily with reference to the intentions behind its formulation. An *heuristic* process, on the other hand, may suspend at runtime or arrive at a point where no choice is made between different alternative solutions. It is sometimes described as leaning on an algorithm that rests on "trial and error", while in other cases it is attributed to a human mind that systematically works through a specific task. At the other extreme end we have what may be called intuitive processes that elude the possibility of any kind of precise descriptions.

[3] in LIS literature, the term classification scheme is often reserved for the application of schemes that adopt a particular notation. However, a classification scheme, as well as an indexing language, makes up what is often termed a controlled vocabulary, and the differences between an indexing language and a classification scheme

tion schemes or indexing languages is used in the same way as for the organisation of document collections in libraries.

Besides a topical analysis, it is possible to analyse document contents with respect to the functions of the documents. There is, for instance, a difference between a bibliography and a research report that is most importantly understood as differences in terms of function — in terms of what is to be accomplished with the documents. Library classification schemes usually incorporate several elements that relate to the functions of documents, rather than their topical contents. Bibliographies, for instance, may be grouped together, as may encyclopediae and literary works of fiction. This is reasonable, as a bibliography mainly has the function of directing the reader to other documents and a research report performs the function of documenting research efforts and results; which do restrict the ways in which the documents can be used. It is assumed in this thesis that these differences express different genres. The words ’bibliographies’ and ’research reports’ are words that usually denote classes of documents primarily formed because of similar aims for their production. But bibliographies also belong to a genre, a set of actions and events, in which the description and enumeration of documents is a common trait. This social conception of the notion of genre, and the physical objects circumscribed by human activities, are two important aspects on genre that have not received as much attention within LIS as has topical content. In fact, with a few exceptions, the notion of genre in LIS is highly undertheorized, which is one reason why the work presented in this thesis might contribute to LIS research.

This task of text classification, which proceeds from an estimation of to which communicative contexts items of a collection of documents adhere, is to be referred to as a task of *document genre classification*. A very important fact when it comes to algorithmic approaches to document genre classification is that algorithms have to proceed from observable data. It is the particular configuration of actions and events from which a genre arises that is of greatest interest; but it

are less relevant here. Lancaster (1998, p. 15ff) points to the often confusing distinctions within LIS terminology with respect to terms such as indexing, classification and subject cataloging.

comes handy that the documents themselves, by virtue of their forms, express typical patterns. A bibliography is well recognized by being a listing of bibliographical references (and not seldom containing the word bibliography). In other words, a human eye may recognize a bibliography because it is highly typified by the form conventions of a genre. Thereby, if it is possible to model those aspects of form that guide a human mind in recognizing an artefact within a genre, algorithms that may assist in such a determination may be formulated.

This typification can be observed at different levels. First, the use of a natural language is normally adapted to the different situations and target communities of the genre. For instance, the first person pronoun “I” is rare in many scientific genres, while common in personal communication. Second, the layout and logical structure of different text elements signal conventions of an extra-linguistic kind that similarly arise from the genre. For instance, a newspaper often has its text arranged in several page columns, while a typical university textbook does not. The linguistic patterns recognizable within the artefacts of a genre necessitate linguistic knowledge that can be used in order to map intrinsic properties of a document to the extrinsic property of genre adherence. The typified layout and structure require a different kind of interdisciplinary knowledge. It may concern the application of hypertext linking or the explicit encoding of different textual elements in order to have them appear in a particular way on the screen or on paper.

When we consider the first typification above, the problem of this thesis is located at the intersection of not only LIS and computer science, but also of linguistics. Parts of various linguistic subdisciplines such as text linguistics, corpus linguistics, sociolinguistics, systemic-functional linguistics, and in particular computational linguistics, thus all have relevance for the problem of text classification according to genre adherence.

Document genre classification is a multidisciplinary task that has attained some interest mainly within LIS, computational linguistics and computer science, while it has only been a computational problem for the two last domains. As a computational problem, document genre classification is a question of mainly three things:

First, how can, for a given collection of documents, a space of genres — a document genre classification scheme — be defined? Second, how can a set of documents be classified with minimal human intervention, or, put somewhat differently, what computational model performs best? Third, which linguistic and extra-linguistic document features[4] have to be considered in order for a document genre classification to be as accurate as possible? These are three very general questions that have been addressed before in different research contexts. They would require a much too wide study to be adequately addressed in full depth within this work, and, as will be shown, neither of these questions seems to have a definite answer.

However, since this problem is a relatively new area of study, there are certain more specific aspects of these questions that have not been particularly well explored. For instance, there remain several kinds of features that have only tentatively been examined this far, and the effects of different granularities and cardinalities of genre spaces are not well known.

1.2 Motivation

The motivation for this work has several different faces depending on from which perspective it is looked upon, but its main motivation is to contribute to the development of LIS in the following way.

Document classification as a library activity still relies on the principles for cataloguing that were presented in the late 19th century by Charles Ammi Cutter, the original designer of the classification scheme used by the Library of Congress. One of these principles stated that the catalogue should “show what the library has ... in a given kind of literature”. Most advanced information systems elaborated for the retrieval of bibliographic information provide a way to restrict a specific query to a certain kind of literature. For instance, the interface for the database LISA provided by ProQuest offers the pos-

[4] A document feature is in this work understood to be not only a property of a document but a property whose value is supposed to differ between different documents in a significant way.

sibility to restrict a search to “conference reports”, “book reviews”, or “literature reviews”. These three labels are names given to classes of documents that are grouped together because they share a certain purpose. In the terminology of the database in question, they are referred to as different “publication types”. It is, however, tempting to say that they are names of classes of documents of importance within the same genre, because the documents are generally aimed at a certain audience in need of documents that are published with a particular situation in mind. However, in this case there are just a few named kinds of documents, and one needs to understand what type of documents they refer to. The “kind”-ness of documents are far from equally well exploited as the “about”-ness of documents in such bibliographic systems. The exploitation of this “kind”-ness and its relation to documentary practices is astonishingly scarce within LIS as a whole. The property of genre adherence is to a large extent ignored, at least in explicit terms, although there certainly are exceptions. Genre adherence relates the documents to the practices in which they are embedded, which has recently become more and more recognized as part of what determines their usefulness and cannot be ignored, but we still do not know exactly what users look for when identifying the kind, type, or genre of a document.

Bernd Frohmann (2004, p. 387) expresses firmly how documentary practices are of outmost importance for information access:

> ...the informativeness of a document depends on certain kinds of practices with it, and because information emerges as an effect of such practices, documentary practices are ontologically primary to information.

This work represents an ambition to respect the importance of documentary practices in systems for information access, and tries to investigate this aspect with special regard to document classification. Its results may promote further exploitation of real world applications that incorporate views on a document collection that reflect its genre variation and can be used to support topical search systems.

In addition, the more specific motivation rests on a need for more explorative attempts within document genre classification to investigate different kinds of features and genre granularities, mentioned at the end of the preceding section.

1.3 Aims and contributions

The problem of this thesis is a multidisciplinary problem of academic study, still in its infancy. As such it suffers from a lack of consensus with respect to different concepts and how the different problems are best approached.

LIS has focused on the design of classification schemes where the notion of genre has not been particularly well articulated. Computer science has mainly been interested in the development and improvement of algorithms, while linguists have mainly been occupied with the study of language use within restricted domains.

If genre is taken in its sense of social action, it must be asked whether this is a way of understanding the word genre that is common within LIS, linguistics, and related application oriented disciplines, such as computational linguistics and information retrieval, and, in addition, if it is compatible with how it is understood within these disciplines. It must also be asked whether what is understood about genre as social action is something that is at all considered within these disciplines. A first list of research questions for this work is thus the following.

1. **How is genre conceptualized within LIS, linguistics and related disciplines, especially with respect to classification purposes?**

Within the application oriented areas, where it is assumed that classification according to genre is being done, it can be asked how the three questions on defining a genre space, applying a classification model[5], and deriving features that correlate with genres, are approached. This leads to two more research questions.

2. **How are different applications of document genre classification realized?**

3. **To what extent do classification applications comply with**

[5] The meaning of the expression 'classification model' will be elaborated on and defined in Chapter 3

an understanding of genre as social action?

The answers to these three questions all arise from the literature. They are, so to speak, posed in order to sketch a framework for a more concrete contribution to the knowledge of how document genre classification can be successfully accomplished or not. New questions have to be posed that are not sufficiently tackled in experimental research, so the three general questions on defining genre spaces, applying classification models, and deriving features will form the foundation for a set of experimental research questions that relate to a fourth and last general research question.

4. **How do different definitions of genre spaces, classification models, and document features influence document genre classification?**

This final, and more compelling question, thus has to be refined in terms of a few experimental questions, which are presented in Part II of this work as they depend on some constraints defined by how an experimental setup can be configured.

1.4 Outline

As this work has two faces, a theoretical and an experimental one, the main body of this thesis has consequently been divided into two parts: *Part I* contains an investigation of the multidisciplinary status of genre and classification and arrives at expressing a particular stance towards document classification according to genres. *Part II* shows how this can be realized and examines to what extent it can be successfully applied. The thesis ends with *Part III*, where some conclusions from the experimental part are drawn, together with a summarizing discussion on the outcome of this work and possible directions for further research.

Part I starts with an investigation of the notion of genre and related concepts and how it has thus far been approached, first within LIS and library classification in particular, then within certain areas of linguistics, and finally how modern text technology[6] expresses similar

[6] The term “text technology” denotes all the principles and techniques that assist

concepts. This chapter should be seen as defining how genre is understood in this work and constitutes as a whole, an answer to reasearch question number one. Chapter 3 introduces a formal definition of classification in order to clarify and define the main issues of this work. It then sketches out the main problems related to the implementation of classification in general. This chapter also tries to give a synthesized view on both human classification theory and practices and their algorithmic counterparts. Chapter 4 gives a concentrated overview of previous research related to the identification or classification of texts based on any kind of genre aspect. Chapters 3 and 4 together answer research questions two and three, and the implications of these answers for experimental issues are summarized in Section 5.1.

Part II starts by presenting the framework for the experiments performed in chapters 6 and 7, including the empirical data used, the classification models applied, and the sets of document properties that are used. Given this framework, a set of experimental questions that arises from the fourth research question closes Chapter 5. Chapters 6 and 7 report on the actual experiments performed, where the results are briefly commented in close connection to the presentation of each experiment. Both chapters end with a short overview of the experiments of each chapter.

Part III contains two chapters, where the first discusses the conclusions that can be drawn from this work with respect to the research questions. The final chapter attempts to determine what we do not know but need to learn in order to proceed with research on genre classification.

in the production of texts.

Part I

Towards a Multidisciplinary Theory of Document Genre Classification

Chapter 2

Genres and text typologies

The understanding of genre briefly explained in the introduction (page 1) conforms to how genre has been treated within the so called *new genre theory* (see e.g. Freedman & Medway (1994)). As such it differs more or less from how it is generally understood in several other disciplinary areas and in common English usage. This chapter will try to clarify these differences through an investigation of how genre has been approached within the domains of LIS, linguistics and text technology, in that order.

More specifically, this chapter is organized in the following way. Section 2.1.1 identifies a distinction between topicality and non-topicality in library classification schemes, since classification schemes also incorporate aspects in between the notions of topic and genre. This is further elaborated in Section 2.1.2, where pure non-topical designators are investigated along with what has been referred to as "form subdivisions" in library classification. In Section 2.1.3, an account of how genre has been studied in LIS is given, with special attention in Section 2.1.4 to the emergent document theory trend of LIS. The linguistic perspectives on genre and its related notions "text types" and "register" are reviewed in Section 2.2, while Section 2.3 is devoted to text technology. Sections 2.4 and 2.5 summarize what can be stated on genre and its recognizability.

2.1 Library perspectives: documentary practices

The classification schemes used today by libraries that organise general document collections (i.e. that are not restricted to narrow domains), such as the Dewey Decimal Classification system (DDC) or the Universal Decimal Classification system (UDC), are usually said to consist of a structure of labels that refer to a semantic hierarchical structure of topics, concepts or subjects. In the words of the renowned "classificationist" Ranganathan, the act of classification itself is "the process of translation of the name of a specific subject from a natural language to a classificatory language" (Ranganathan, 1994, p. 31). Ingetraut Dahlberg, another influential classificationist, states that "the elements" of classification schemes are "concepts or representations of concepts" (Dahlberg, 1978, p. 9).[1] One could thereby conclude that when a concept is chosen for a class, the concept in question refers to something which should be shared by all documents of that class, and that this concept is treated by the documents. However, taking the label '011' in the DDC as an example, it refers to a class of documents that has the common feature that they are bibliographies and not about bibliographies. There is an important difference between bibliographies as a topic and as kinds of documents, where the latter aspect is often referred to as a matter of form but is, essentially, not really that simple, as will be claimed below.

2.1.1 Subject matter versus form

Form is often contrasted with content in classification practices. It is obvious that e.g. general bibliographies are not given a designated class because of their topical properties, since general bibliographies are not *about* something particular. Bibliographies are thus said to be classified according to form. This may be misleading. It is not a case of suddenly having a group of documents without content. All

[1] Note that this quotation also expresses a shift of focus from the division of a collection of documents to the translation of a subject analysis, which will be further discussed in Section 3.3 of the next chapter.

documents have content and form, it is just that the meaning and potential use of bibliographies are determined not by the topics treated, but by their intended use, or what the bibliographies may do for the user who knows how to handle them. It seems a misguiding simplification to equate the content of a document solely with what a document is about. In many cases, what seems to matter the most is what a document is about — but that is far from always the case. It seems a comparable simplification to state that for some documents form is what matters the most; it is only that in the process of classification, form is considered the most convenient property to use as a discriminator.

It is not altogether clear what is meant with form in bibliographic practices.[2] The word "form" denotes many different aspects of documents. From the perspective of Wilson & Robinson (1990, p. 39), bibliograpies are distinguished by their non-discursive character, photographs by being non-linguistic, and manuals by not being intended for consecutive reading — binary characterizations that are rather different from each other. Form in bibliographic practices is a manifold notion and a generic denominator for non-topical aspects on documents, rather than something distinct. It must be admitted that all documents have form and content, but not all documents have easily determined topics.

Let us start here with an examination of how topic is contrasted with other document properties in bibliographic practices. The term "topic" is often used interchangeably with the term "subject" in LIS in general. However, subject seems to be preferred by those who design and revise classification schemes, and taken to be something more general than topic, while topic is preferred in information retrieval research, especially when connected to TREC experiments, where it occupies a core position together with the notion of relevance.[3] The

[2] Bibliographic practices are understood as all those activities that aim at analyzing or describing a document in some way. It is an extensive area of practices which includes both enumerative and analytical bibliography, where the former is mainly aimed at enumerating what has been published within a certain domain or time span, while in the latter studies can partly be characterized as more archaelogical. (Cf. Dahlström, 2006)

[3] The Text REtrieval Conferences can be described as an ongoing contest between

terms will be used interchangeably in the following, respecting the wordings in the texts referred to, but this is not to imply any sharp distinction between the meaning of the two words. Subject is defined in ISO standard 5963:1985 (*Documentation — Methods for examining documents, determining their subjects, and selecting index terms*) as “any concept or combination of concepts representing a theme in a document”, whereas “concept” refers to “a unit of thought”. This definition introduces the notion of theme, which is also used in place of topic. But let us first illustrate the distinction between topical and non-topical statements with two simple statements.

```
This book is about bibliography
```

Example 2.1.1

```
This book is a bibliography
```

Example 2.1.2

The first statement is a statement on the subject, while the second one is not. Such a simple linguistic test should in many cases be enough to determine whether what can be said about a document is a characterization of its subject. If it is appropriate to say that a document is about X, then X is a subject denominator. The words ’subject’ and ’topic’ are in fact sometimes substituted by the word ’aboutness’ in LIS (see, for instance, Hutchins, 1978). However, sometimes we run into trouble with the linguistic test. Consider a timetable for the local bus company, or a directory of telephone numbers. These are examples of a timetable and a telephone directory. It is not hard to say what they are or are intended to do, but it would be rather awkward to say that they are about bus traffic and telephones *in the same way* as the annual report of the local bus company or the telephone company. Still, it is possible to say that the telephone directory is about telephones, or telephone numbers and people.

In some cases, thus, the linguistic test is not enough. Consider now a thesis that treats the development of the socialist movement in

researchers concerned with different kinds of algorithmic applications.

Russia, with obvious historical perspectives. Is this book about history? In some sense we can probably answer yes, but it would be equally possible to answer no, depending on our linguistic intuition. Langridge (1989) would probably refer to such an example as being a case of a book having history as its "form of knowledge", whereas socialism would be the topic. As a thesis, the document has to be produced within the context of some academic discipline, most likely that of history. History would, from Langridge's perspective, be seen as a way of "looking at the world" (p. 31). This is fairly consistent with Mills & Broughton (1977, p. 36) in their explanation of form of knowledge: "the concepts and methods of enquiry". Determination of the form of knowledge and the topic are both part of subject analysis and, Langridge (1989, p. 45) states, "exhaust the idea of subject matter in documents". However, for Langridge, discipline attribution is not part of subject analysis, although this is explicitly stated as the most important principle for subdivision in the DDC: "the parts of the Classification are arranged by discipline, not by subject" (Comaromi et al., 1989, p. xxvi). If we consider another example, a typical introductory textbook for university studies in history, it would in fact be hard to find any other term than history that is encompassing enough to describe what it is about. No one would probably object to say that it is about history, although it is not about history in the same way as in the example of the history of socialism in Russia. Clearly, there is a difference here that may be explained as related to differences in conceptualisations and methods.

Although bibliographic classification schemes are often seen as mirroring classical subdivisions of human knowledge, these subdividing principles seem to reflect the division of academic disciplines as well. When we talk about studying a certain subject, such as history or chemistry, this does not mean exactly the same as when we say that the topic of our discussion is a certain subject matter. The former sense is tied to an institution, to certain communities of academic practice, whereas the latter does not have to be. The distinction between topics and forms of knowledge seems to mirror differences with respect to degrees of dependency on academic communities of practice. Considering the heritage of classification schemes as scientific knowledge classification, as claimed by Miksa (1992) and Hansson (1999), it is not surprising to find instances of both topical designators and designators of academic disciplines in classification schemes. However, far from all documents in most general collections are scholarly works, and may thus be inappropriate to relate to academic communities. A book on car repairs, for instance, is related to certain practices, and it is possible to see forms of knowledge as intimately related to practices in general, although not necessarily to academic practices.

With the first example above (Russian history), it could be reasonable to say that the topic is ’socialism’, or whatever term is preferred according to the controlled vocabulary chosen, and that the academic discipline or community of discourse and practices in which it has been authored is ’history’. The second example above, the textbook in history, may then be similarly designated as a book within the domain of history studies. The topic is, strictly speaking, not history, but possibly the domain of history studies, if the book makes explicit claims of characterizing the study of history as an academic discipline. Thus, it is now apparent that in addition to topic (and form), bibliographic classification is also concerned with something in between topicality and characteristics of form.

Besides forms of knowledge, Langridge states, there “remain a number of very important characteristics requiring identification which have always been treated as part of the process of subject analysis” (1989, p. 45) The “important characteristics” that Langridge refers to as not strictly related to topic or ’form of knowledge’ are, for instance, the viewpoint from which a piece of text is written and the level of expertise required to read it. He groups these characteristics under the heading “forms of writing”. Forms of writing is a convenient addition to the classification schemes, because it makes it possible to classify material that is not topical in any obvious way. According to Miksa (1992, p. 110), several kinds of non-topical additions to the schemes stem from the beginning of the 20th century, when document retrieval gradually became the primary purpose for library classification. Sukiasyan (1998, p. 75) places it even earlier in time, in 1879, with Cutter’s supplement to his “Expansive Classification”. The so called “form subdivisions” have since then been the object for classificationists' discussions and form subdivision has turned out to be a notion of several meanings. In fact, it seems to be more of a generic term for non-topical classificatory aspects (see for instance, Wilson & Robinson, 1990, Taylor, 1999, pp. 142-143).

However, if it is appropriate to say that a document is an X, then X is a designation of the kind of document, a kind which is not topically determined and possibly related to form, because form is that which meets the eye before any deeper interpretation takes place (cf. Wilson & Robinson, 1990, p. 37). All documents will in some sense be appropriately described as being something that is not at all topical and having a certain characteristic form. There is always one or more form-aspects on documents, although several forms of documents are not the subject of classification in libraries. However, as with the example of bibliographies, it is not really their form that matters, but something else. Form is only the means whereby the identification of a bibliography is easily done.

2.1.2 Form subdivisions in classification schemes

Having associated apparent non-topicality in classification schemes with what is commonly referred to as "form subdivisions", and in some way related to documentary practices, provides us with a clue to how non-topicality is understood in bibliographic classification practices. It still remains rather vague, though, and there is a need to look at what is really implied with form subdivisions.

Wilson & Robinson (1990, pp. 39-40) enumerate six different groups of form subdivisions found in a classification guide. This enumeration represents a step-by-step exclusion of documents based on modes of perceptional access and intended ways of reading. Form subdivision proceeds by first eliminating non-verbal works, then formatted data of a non-discursive character (including e.g. bibliographies), verbal expressions that are not expected to be accessed in a sequential way, fictional works, composite works, and finally moves on to (nonfictional) genre subdivisions. Genre subdivisions are exemplified with "case studies, comparative studies, comic history, interviews" but "share no common character other than in one way or another relating to the kind of writing that can be expected . . . ” Wilson & Robinson are particularly occupied with the idea that there are no such things as documents that do not lend themselves to form subdivision. Description of genre is applicable to almost any document and is important because “genre or kind is the idea of a range of conventional procedure that guides both the performance of producers . . . and the expectation of users” (p. 42). Their observation of the communicative role of genre is consistent with a general idea of genre and the understanding of genre in this work.

Taylor (1999, pp. 142-143), with reference to the approved form definition of the American Library Association, enumerates five types ranging from the physical character of documents (media type and type of expression, such as photographic material) to literary genres (e.g. drama). Here, again, the word genre is encountered, although in the sense of literary genres. The aspects of form that distinguish novels from poetry and drama are in LIS and library practices often referred to as *genre* characteristics, for instance, in the LIS encyclopedia of Reitz (2004). Otherwise the word genre is mostly ignored in most LIS encyclopedias. Feather & Sturges (1997), Kent (2003), Drake (2003), for instance, have no entry on genre, not even in the indices. Form, with respect to literary genres is not the same as form in the case of bibliographies or, for that matter, in the case of media types. A recent exception of ignorance, which also witnesses an increased interest in genre theory within LIS, is the entries on “Genre Theory and Research” and “Internet Genres” in the third edition of *Encyclopedia of Information and Library Sciences* (Schryer, 2010, Crowston, 2010). The first entry, however, does not elaborate on the notion of genre with respect to information seeking and classification, whilst that is the case for the second entry.

The aspects of function, or intended use, that distinguish multilingual dictionaries from bibliographies and term dictionaries are sometimes referred to as differences with respect to *document type*. There are other terms in bibliographic practices in frequent use that signify similar aspects that have little or nothing to do with topic, such as publication and object type, and which falls into the categories of ’form subdivisions'.[4]

In bibliographic description in general, as realized in contemporary cataloging practices governed by the scheme of the MARC21 format, it is possible to label a document representation with codes that signify, for instance, "the nature of contents" (e.g. if a document is a PhD thesis or a legal article) and "target audience" (Library of Congress, 2004) at certain positions of the fixed field 008. However, the possible codes designated are mixed with codes referring to categories other than genres, such as "sound".

So, besides topic and form of knowledge we now see that there is a wealth of non-topical document aspects that are given attention in bibliographic description and classification. Many of these relate more or less to documentary practices — what the documents do and how they are used. "Target audience" is nothing but a particular kind of explicit specification of the community to which the documentary act is directed, and "the nature of contents" often relates to the purpose(s) of a document.

In the list "basic genre terms for cultural heritage materials" developed for the American Memory project we likewise find genre designators mixed with such designators as "books" and "clippings".

All bibliographic element types can in fact be used in classification tasks. It is, for instance, common in library shelving practices to group some 'form subdivisions' (e.g. journals and reference works) separately, either completely separate from the rest of the library collection, or separate within a top level class.

Genre, in its explicit sense of social action, is only rarely explicitly reflected in classification and cataloging practices. Genre is often counted among the many form aspects, but in contrast to the varying requirements on modes of perception and reading that Wilson and Robinson refer to, genre is determined by more encompassing factors, related to other dimensions of the use of documents and their sociocultural context. It seems that in library practices, the focus is on form

[4] Crowston & Kwasnik (2003) seem to regard document type as a generic term for genre, publication type and similar terms. See also the discussion provided by Svenonius (2000, p. 113) on the distinction between different non-topical aspects of documents.

rather than on what the particular form expresses, simply because a genre is often recognizable by artefactual form. Genre cannot be reduced to the form of its artefacts, if genre is understood in a social sense. An often cited explanation from the systemic functional school is that "Genre are how things get done, when language is used to accomplished them" (Martin, James R., cited in, for instance, Swales, 1990, p. 40). When genre is understood in this way, as socially motivated action, it contrasts sharply with how the word is understood as denoting literary or artistic style. The difference between, e.g., a crime novel and a romance relates more to narrative topic than to communicative purposes, and should therefore not be confused with (non-fictional) genre. In fact, within LIS, genres are understood mostly as fictional categories. However, there are some exceptions in LIS that will be referred to in the following.

2.1.3 Explicit genre perspectives in LIS

Topical aspects have been given most attention in LIS, rather than "the way information is packaged", as Svenonius (2000) expresses it. Although this is true, the packaging is not ignored, as we have already seen. The packaging of information "determines its usefulness", she states, and seems at first glance to agree with the quotation from Frohmann at page 8 in this work. However, Svenonius treats these ways of packaging information as "physical and material attributes" and, scarcely related to social action. She includes them under the heading "document languages", along with "publication attributes" and "access attributes", to signify that these descriptive elements provide access to the embodiment of information as opposed to conveying information contents (Svenonius, 2000, Chapter 7).

It is also in this way that Vaughan & Dillon (2006) explicitly express their interest in genre, albeit mainly from the perspective of cognitive psychology. They have performed a user study on how "information space design" influences comprehension, usability and navigation, and found that a genre-conforming design was significantly more effective (cf. how Toms et al. (1999) show that the visual structure conveys genres). Thus, user expectation is claimed to be of outmost importance and it seems, not surprisingly, that innovative design has to be carefully reconsidered so as not to violate user expectation. However, genre is not explicitly defined as a social notion in this investigation, and even though part of their investigation is intended to determine what users imply with a genre-conforming design, it lacks generalizable results with respect to a social notion of genre.

Crowston & Williams (2000), Beghtol (2001), Toms (2001), Kwasnik et al. (2001), and Rosso (2005) are among the other exceptions within LIS that show an interest in genre as an explicit social phenomenon. One of the more in-depth attempts within LIS to study the phenomenon of genre with respect to bibliographic classification is an attempt to apply the notion of facets, derived from Ranganathan's ideas of faceted classification, to the elaboration of a classification scheme for web genres. Crowston & Kwasnik (2004) attempt to identify what "clues do people use to identify genre when engaged in information-access activities?" and group these into what they call "facets". Among the facets they count are e.g. structure, language level, graphics, and (document) length. The "clues" they have identified range from fairly specific ("more than 5 pages long", ".edu in URL") to more vague and open-ended clues ("artistic layout", "particular style of photos"). Even though they explicitly adopt a social notion of genre borrowed from communication studies, their focus seems to remain one of form rather than of socially based function. Crowston and Kwasnik claim that they have chosen a bottom-up approach as opposed to a usual top-down approach, in asking questions about how the user perceives and understands different genres. This may be true, but they do ask these questions in order to establish a classification scheme that seems to foster a top-down approach, i.e. assuming a stable genre space to which documents have to be mapped, or in other words, a fixed set of categories to which documents have to be assigned.

Rosso (2005) sees genre as a "folk typology" and takes for granted that a class of documents that is not recognized as belonging to a genre is not to be considered a genre, at least not with respect to that group of users. Even though he explicitly adopts the view of genre as a conflation of form, purpose and content, his view is very strong on the point of user recognition. This, however, seems fairly natural as he appears to consider classification along genre dimensions mainly as a support for querying[5], in which case genres that are not consciously known and given names are fairly useless. This does not have to be the case for browsing, if documents can be visualized in groups. Similar to Crowston & Kwasnik, Rosso's aim is to establish a genre space, based on a systematic user-centred work with involved informants of different kinds and different sizes.

In 1997, Anders Ørom wrote an article in the Danish library journal *Biblioteksarbejde* (1997), which marks a start of interest in genre within Nordic LIS research. Ørom's view is that a genre is characterized as a conflation of functionality, the use of language, its mode of presentation and the author's position within the text. (1997, p. 8) Ørom uses Roman Jacobson's model of communication to elaborate on the use of language in genres, where communicative functions of referential, emotive, phatic, connative, poetic and metalinguistic character determine the configuration of a certain genre. In addition, he puts forth the idea that genres are connected to either institutional practices or to an open community. Some genres are intimately tied to e.g. academic activities, while others are directly aimed at a common public, which is the case with newspaper articles. As a theoretical framework Ørom's article is interesting, but it fails to show more than this. There is no detailed attempt to propose its application within knowledge organisation.

In Denmark, the "epistemological lifeboat" (Hjørland & Nicolaisen, 2006), said to be an introduction to the "philosophy of science from the point of view of Library and Information Science", includes a section on genre by Jack Andersen. Andersen has paid special attention to the notion of genre as it is understood within the North American school of rhetorical studies, of which the article by Carolyn Miller (1994), referred to in the introduction (page 1), marks a starting point. In his thesis Andersen uses this new genre theory as more of a theoretical framework to study the relationship between knowledge

[5] Querying takes place when a user input keywords or phrase to be processed by a database engine. Section 3.5 elaborates further on different modes of access to document collections.

organisation and social organization, to “illustrate how activities and practices based on the use of documents get typified with regard to the maintenance of a given social organization“ (Andersen, 2004, p. 22).

In this respect, Andersen makes a similar use of the concept as in the often cited works of Orlikowski & Yates (1994), where genre is defined as “a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form”. They use it, as well as Honkaranta (2003) and others, for the study of organizational communication. None of the latter works are from within LIS, but signify a particular kind of analytical use for the notion of genre that is considered fruitful but has less to do with bibliographic classification.

It should also be mentioned that a dominant trend in some parts of LIS is to apply discourse analysis as inspired by e.g. Michel Foucault, Norman Fairclough, Charles Laclau, and Chantal Mouffe. However, despite the strong focus on language use in discourse analysis, the objectives of these studies are directed more towards the study of power relationships and/or information user behaviour within communities of practice than towards its application for bibliographic classification or the delineation of artefactual typification. The connection between LIS and documents as socially situated artefacts is probably the most explicitin the trend towards document theory, which is the focus for the next section.

2.1.4 The document theory trend

Since 2003, there has been an annual meeting, starting in Berkeley, California, termed *The Annual Meeting of the Document Academy*. These meetings, in the form of small interdisciplinary conferences, were initiated by the School of Information Management and Systems at the University of California, Berkeley, and the Department of Documentation Studies at the University of Tromsø. They are focused on documentation issues and documents are forefronted as objects of and for social action. LIS representatives are in majority and the meetings can be said to mark an ongoing trend in LIS with a shift of discourse from information towards the materiality of documents.

As one of the papers from the first meeting points out, the transparency given to documents in the study of information is historically contingent (Levy, 2003b). It is a result of abstraction, where content is disattached from its material embodiment. Levy draws on the works by Ivan Illich and sees a parallel to the notion of information in how Illich describes how the conceptions of books and reading changed in medieval times. In the 12th century, Illich writes, the *text* "had begun to float above the page" (1993, p. 118) and the text

> could now be seen as something distinct from the book. . . The text, rather than the book, became the object in which thought is gathered and mirrored. (p. 119)

Before that time, reading as well as writing took place as oral (and social) activities, and the book functioned as a kind of mnemonic device in these rituals. Levy (2003a), with a background in computer science and calligraphy, has been particularly interested in what documents do and how they are given delegation. Documents are "surrogates for people" (Levy, 2000, p. 24) and the technological innovations adapted for documentation practices pose special problems. This, we can assume, may be one of the reasons for a renewed focus on documentation instead of information within LIS. Michael Buckland is one of those who have tried to pinpoint different understandings of information within information science. Buckland (1991, p. 39ff) particularly observes that interpretations of the word "information" are sometimes intimately dependent on belief and sometimes related to truthness and falseness. Belief, truthness and falseness are of course often of importance for the impact of documents, but a more striking impact of documents may be, for instance, their ability to uphold the existence of communities. This was certainly the case with the fanzines of the late 1970s and early 1980s, without which the postpunk music scene had been something completely different. This is emphasized and further exemplified by Brown & Duguid (1996).

When content is disattached from documents, it is often mapped to the human mind and discussed in terms of information. Just as topics are inferred from documents and mapped to the conceptual structures of classification schemes, which gives one and only one level of document representation. The so called cognitive viewpoint in LIS, scrutinized by Frohmann (1992) and criticized by Hjørland (1998), expresses a shift from the study of documents to what is possibly going on inside the human mind. An intermediate approach in LIS is the now frequent shift to some form of social constructivist approach, common within user studies. In these latter approaches, however, most attention is given to users' thoughts and experiences, rather than observing how material objects become part of social life. It is users and their social interaction with something abstract, information, that are in focus.

The document theory trend in LIS, if we may call it that, could in part be described as a reaction against the backgrounding of documents as material objects and an emphasis on their dependency on sociotechnical conditions. In many cases it can be described as trying to bridge the "great divide between social sciences and computer science" through "technically informed social analysis", as it is put in an introduction by Peterson Bishop et al. (2003) with respect to the study of digital libraries.

In explicit terms, genre is not foregrounded in this trend, but the view on documentary practices is particularly reminiscent of an understanding of documentation as an inherently sociocultural phenomenon, susceptible to genre conventions. For Frohmann (2004, p. 387), the documentary practices in relation to their informativeness, in Frohmann's own words, concern 1) their materiality, 2) their institutional sites, 3) the ways in which they are socially disciplined, and 4) their historical contingency. Levy (2003a, p. 33) says about a document, that 1) it was "created at a certain time and place", 2) "comes out of a certain community", and 3) "exemplifies a certain style and design aesthetic", which conforms with the genre triple of communicative situation, community of discourse and artefactual form.

A recent thesis by Francke (2008) conforms with this trend in focusing the practices around open access journal initiatives and the roles which the journals play on the scientific publishing arena, with a particular interest in how document and information architectures are realized. This work may to some extent be seen as a follow-up on the call for the study of document architectures within LIS proposed by

Dahlström & Gunnarsson (1999). The relationship between matter (or text technology) and context, often otherwise ignored in LIS, is in this work forefronted, as is also the case in the works of Dahlström (see e.g. Dahlström, 2002b,a, 2006).

As the label indicates, the document theory trend is mainly theoretical, with some exceptions. It proposes alternative perspectives on the scope of objectives for LIS research and is in a sense opposed to the traditional delimitations of LIS. Although, it must be admitted, the explicit precedence given to discipline rather than subject matter in the guidelines to the DDC scheme shows that the use of documents is given attention: "works that are used together are found together" (Comaromi et al., 1989, p. xxx) However, with its focus on the "social life of documents" (Brown & Duguid, 1996), the document theory trend rejects the positivistic or mentalist approaches that LIS has been criticized for (cf. Hjørland, 2002, Frohmann, 1992). The practical implications for library practices, however, do not seem clear. Do we expect classification, indexing and cataloguing to be significantly different as a consequence? Are we to design information systems differently? Perhaps we may say that the importance given to provenance by archivists, which implies due attention to a document's relation with community practices, should be more reflected upon in library practices. Moreover, the notion of form subdivisions, i.e. genre in particular, may deserve increased attention in classification. In addition, the methodological approach of bibliometrics, with its focus on statistical measures of e.g. co-citation is inherently concerned with studies of the use of documents rather than with studies of the contents of documents. However, bibliometrics is mostly applied to the strengthening of topical retrieval and concerned almost exclusively with scholarly documents. Thus, all methodological approaches that deal with the use of documents, including bibliometrics, scientometrics and webometrics are in some sense theoretically conformant with the document theory trend as described above.

2.1.5 Concluding remarks on genre and LIS

In explicit terms, genre is seldom recognized in LIS although several properties described as related to form (in library practices) seem to reflect aspects on genre. However, an increased interest in sociotechnical and sociocultural aspects on documents and libraries has to some extent drawn attention to the notion of genre, by means of an increased interest in the interplay between context and technology, but its function as a classificatory principle, especially for bibliographic practices, is largely undertheorized and unexploited.

When genre is given attention as a classificatory principle within LIS, it is mostly regarded as a problem of enumerating or specifying sets of genres and as a form-property of a document, not as a property of its context. Even so, these efforts are not to be ignored when they embed intitiatives that put users' perspectives in relation to genre.

2.2 Linguistic perspectives: Language use as action, genre as action

Linguistics is, compared to LIS, a more established discipline and field of study. Linguistics may be applied for fairly different tasks. Let us consider a few. Linguistic theory and its methods can be used in order to describe linguistic competence for language learning. This has been a prevailing objective for some time, especially professed by e.g. Noam Chomsky. It can also be used to describe the actual use of language, which has been given an increasing interest since it became possible to computationally process large quantities of language data (see for instance McEnery & Wilson, 2001). In fact, it is the analysis of actual language use, as opposed to linguistic competence, that makes it possible to predict language use, which in its turn is a requirement for many different applications, such as information retrieval systems, dialog systems, and question-answering systems. At this point, the objectives of linguistics and LIS converge. In order to give appropriate access to large quantities of text, models have to be developed that make it possible to use the linguistic contents in order to predict whether pieces of texts are relevant or not with respect to a particular task or topic.

From one perspective, linguistics is concerned with graphemes (or phonemes, for spoken language), morphemes, words, phrases, clauses, and sentences of natural languages. In that respect, documents in the sense larger units of text *and* artefact are somewhat out of scope for some linguists. For instance, the branch of linguistics that is called semantics is occupied with trying to pinpoint the "conventional meaning" of words and clauses (Yule, 1996, p. 114). The expression "conventional meaning" denotes the fact that individual, sociocultural and situational variations are of marginal interest. Nor is the full meaning of a complex utterance at a special occasion in focus. In LIS there is, on the contrary, (ideally) a need to grasp the full meaning of complex systems of utterances, not just the lexical (or conceptual) meaning of its smaller constituents regarded as context free utterances. Interrelations between clauses are considered and put in context of their use in that branch of linguistics which is termed pragmatics, while in the case of text linguistics or sociolinguistics the focus lies on complete systems of textual expressions and their contexts.

Sociolinguistics is the name given to that part of the linguistic discipline that studies "the use of language in contexts of situation" (Hymes, 1974, p. 3), as expressed by one of its earliest proponents. Sociolinguistics focuses on language use that arises out of communicative motivation in social life, and thus constitutes valuable background knowledge for the detection of genre from linguistic expressions.

An early attempt to picture language as something highly dependent on real world situations is the speech act theory formulated by Austin (1975), where a linguistic utterance "is, or is part of, the doing of an action" (p. 5). In his influential 1955 lectures, Austin started out by observing that many utterances did not conform to the common philosophical conception of a statement, they did not constate anything that could be characterized as being true or false. He termed these non-constating utterances *performatives*, because they are used to achieve some particular goal, to perform a job, without which the goal could not be achieved. Common and distinct examples can be found in the context of marriage and baptism ceremonies, where the utterings of certain phrases are the necessary requirements for the fulfillment of the ceremony (where documentation plays a central role). Consider a more subtle example, a research article in which an utterance begins with

```
We define classification as being a human necessity
```

Example 2.2.1

This is obviously different from

```
Classification is a human necessity
```

Example 2.2.2

The phenomenon related in both cases is the task of classification. Classification is the focus, and it would be plausible that if anyone should try to map these utterances with classes of a library classification scheme, they would fall into the same class. However, in the first case the author does not state anything that is to be considered true or false. The author is simply stating in what sense he or she is going to use the word 'classification'. By this utterance is enacted an implication that must be obeyed by any subsequent utterance, if communicative rules of conduct are to be obeyed. In the second case the author (for some reason) describes an innate character of the human being, which can be approached as true or false. The first case clearly illustrates a performative, but what about the second case? In isolation it conforms well to constatives. It says something and is clearly different from the first case. It also states something that is meaningful to judge as being true or false. But does it really not imply some kind of action as well, apart from transferring an idea of classification? Actually, it does.

Considering the syntaxes of these two constructs, it is important to note that the main predication in the first case is towards the pronoun 'we', and in the second case to the noun 'classification'. The finite verbs also pertain to different categories of verbs. It may seem that there are syntactical and lexical clues to the nature of clauses as being constative or performative.

In the end of his lectures, after trying to identify a lexical catalog of verbs and verb forms that mark the occurrence of performatives, Austin sums up by observing that the distinction between performatives and constatives is not as clear as it first appeared to be. Almost any utterance implies an enactment of a speech act, and to interpret the speaker's intentions and act upon them accordingly is to understand the utterance. The difference between constatives and performatives is only a slight shift of balance between *locutionary* and *illocutionary* forces — between what an utterance is saying and what it is intended to do. The *perlocutionary* force denotes what an utterance actually accomplishes.

Language is used in order to accomplish certain tasks, such as to convince a person or a group of people, make someone do something, or simply to establish some kind of human contact. The use of language with respect to certain tasks causes the sociolinguist to focus on *variation*. Variation refers to when the use of language in one instance of a particular situation differs from the use in another instance with identical situational settings. Obviously, language must be used differently to convince someone depending on what we want to that person of. The topic is thus one of the factors that influence variation. However, who is to be convinced and our conceptions of e.g. his or her background knowledge, education and expectations are equally important and also influence variation.

To warn someone can be said to constitute a task for which language is used in rather similar ways on different occasions. In other words, we expect rather small variations from one event to another. Let us consider a situation in which a speaker wants to warn someone that something dangerous is approaching and thinks that the best way to escape is to run as fast as possible. Here, the utterance of the simple imperative `run!` may be a natural choice. The situational factors determine the spatiotemporal extension of and the mode and media for this illocutionary act. It has to be short in order to accomplish its task immediately, and for the same reason it is spoken rather than written.

In other cases, variation may be more apparent. Consider a business transaction between two parties situated far from each other. The mode of communication chosen in this case may be a business letter. Here, the letter is a piece of paper with text put in an envelope on which is inscribed certain required and some optional text. The requirements of the postal services limit the possible appearances of the business letter, and it would in fact be rather appropriate to say that the business letter is to some extent defined by technological factors. It is technologically *typified*. In addition, the contents of the business letter have to respect certain sociocultural conventions that govern the use of language and the structuring of the text on the piece of paper inside the envelope. In addition, being typified by the task and the technological constraints, the language and the organisation of parts of the text are socioculturally typified.

The study of linguistic variation thus focuses on how certain aspects of text and artefact vary whil others do not. Where certain aspects remain constant, and others do not, we may identify varying features that are indicative of how the contextual factors (technology, situation, sociocultural settings) vary. Where the technological factors are constant, such as with a letter, the variation of purpose and addressee is reflected in how the linguistic contents vary. Variation in expression is not arbitrary; it correlates (at least partly) with contextual (and individual) variation.

The typifying process can be analysed in different ways. The systemic-functional school does this by attributing different typifications of language use related to situation — to field, tenor and mode (Halliday & Hasan, 1989). Field is comparable to what happens when language is used, the relationship between the text and the real world. Tenor is the functional and interpersonal role given to the text, the task which is delegated to the text. Mode is the form of expression chosen, in many cases technologically determined. The sociocultural context of text is equally important and referred to as genre (see Section 2.2.3 for a more thorough account of the systemic-functional perspective).

Linguists have otherwise been rather reluctant to use the term genre. It may be, as Swales (1990, p. 38) remarks, that the term is often dismissed because of its history in literary studies. It may also be, as Biber & Ferguson (1994, p. 6) point out, that linguists in general tend to see “written varieties as a literary concern” because when interest is focused on language in use, text is often seen as secondary with respect to spoken language. When linguists use the term genre it is used in quite a different sense and sometimes, as observed by Santini (2004b) and Lee (2001, p. 41), in an inconsistent way. The meanings attributed to the word in linguistics seem to range from those denoting characteristics on the levels of expression in a text (i.e. the artefactual level) to those that deal with the underlying sociocultural and situational configurations related to the use of language and text.

A more frequent term within linguistics is *register*, which in short may be said to denote a conception very similar to that of genre but more focused on the use of language *per se*. In a comprehensive English grammar, ’register’ is said to refer to the “grammatical characteristics of particular kinds of text” (Biber et al., 1999, p. 8). The dictinction between register and genre is one of perspective. Register studies are characterised by a focus on language use, often disregarding many of the technological and extralinguistic properties of documents that are accounted for in some genre studies. There are also several other terms that are used within linguistics to refer to concepts similar to that of genre, such as ’sublanguage’ (Sekine, 1998). A comprehensive overview of such terms within linguistics and literary theory is given by Lee (2001).

2.2.1 Studies of non-fiction in the nordic countries

Recently, we have seen some interdisciplinary research initiatives in the Nordic countries on “non-fictional prose”[6] that relate to both genre and register. These initiatives explicitly claim to cross the borders between the humanities and the social sciences. Unfortunately, text technology is mostly ignored, and the influence of modern technology thus not considered. A few of these initiatives is still worth mentioning. In the 1990s, a Norwegian project called “Norsk sakprosa” resulted in Ottar Grepstads *Det literære skattkammer* (1997), which includes an

[6] The term reads “sakprosa” in Swedish and Norwegian. Terms such as “bruksprosa” and “facklitteratur” are also used to denote certain kinds of non-fictional literature.

attempt to formulate a theory of non-fictional prose and to compile a catalog of non-fictional types of writings. Grepstad's work could appropriately be described as a work of text linguistics and text history. Even though the focus is on texts as linguistic artefacts, detailed linguistic and sociological perspectives are scarce.

From the tradition of text linguistics and text grammars (see Section 2.2.4) Grepstad borrows the idea that non-fictional prose primarily has to be categorized according to the notion of ahistorical *text types*. These text types reflect common communicative purposes and are dependent on text composition, the ordering of text elements, style, orthography, and form of language (Grepstad, 1997, p.504). Consequently, they are not only stylistically determined, but reflect a certain "skrivemåte" (form of writing). The text types are derived from the intentions behind a text and from the position of the author. These are, freely interpreted into English, 1) the argumentative, 2) the expositive, 3) the narrative, 4) the descriptive, 5) the educational, and 6) the normative text types (Grepstad, 1997, pp. 113,164).[7] A text type is often a combination of *text elements* with different typological characteristics, but dominated by one. For Grepstad, the notion of text type is superordinated to that of genre, since genres are conventionalized realizations of text types. Grepstad identifies fourteen genres of non-fictional character — e.g. biographies, topographies, and scientific prose. Non-fiction is seen as a main genre ("hovedsjanger"), on the same level as epics, poetry and drama, and to some extent, as being ahistorical — i.e. stable over time. Each genre includes subgenres with common characteristics, such as theses, reports and articles within the category of scientific prose. The notion of genre, for Grepstad, has many implications, besides denoting conventions for the use of language. He regards genres as social institutions determined by combinations of form and content (Grepstad, 1997, Chapter 5). Genres are thus more determined on the basis of extratextual criteria than are text types.

For Grepstad, the notion of non-fictional prose has a rather restricted scope, and it is not difficult to find examples of non-fictional

[7] Cf. Werlich's text types on p. 41ff in this work.

prose that do not allow themselves to be incorporated into his scheme. Shopping lists, recipes and personal home pages have no obvious place in his catalog. In addition, his theory seems to imply that the existence of a genre name is a prerequisite for recognizing it as a genre.

A Swedish interdisciplinary research initiative is different and dismisses Grepstad's somewhat static genre conception. Here, text type and genre are two distinct phenomena. Text type is a generic term that signifies many different distinguishing principles for classification of non-fiction, including topic, target audience and author (Hellspong & Ledin, 1997, p. 20). Grepstad's text types correspond to the Swedish group's *form of presentation* ("framställningsform"), whose characterization is based on linguistic features solely — primarily how coherence within a certain text is realized. For instance, if text coherence[8] is established by temporal links between clauses and sentences, it is said to exemplify a narrative text, if it is adversatively established it is said to exemplify an argumentative text. Genre, on the other hand, is said to be based on 1) commonly established names of text types, and 2) its social foundation. Genres are to a great extent contextually defined and their relationships with contexts can be analysed along three axes: situationally, intertextually and culturally (Hellspong & Ledin, 1997, p. 24).

This view on genre is interesting because of its stress on intertextuality, which is not otherwise explicitly investigated together with genre. Hellspong & Ledin (1997) distinguish between vertical and horizontal intertextuality. Verticality is established between the text at hand and its implicit dependency on recent texts within the same genre, whereas horizontality is established by more or less explicit references to other texts within the same genre or across genre borders (Hellspong & Ledin, 1997, p. 56). This perspective becomes emphasized by Englund & Ledin (2003, pp. 203ff) where intertextuality is extended beyond intrinsicality and related to sociocultural contingencies. Here, in this last study, the interest in genre shifts to discourse analysis in its Foucauldian sense.

[8] Text coherence denotes the situation where a piece of text is perceived as a unity, as opposed to a set of atomary linguistic expressions.

2.2.2 Genre theory

Maybe one of the most influential views on genre professed in recent years is the one found in what has sometimes been termed the "new genre theory". A paper by Miller (1994), originally published in 1984, marks the starting point for this sociologically influenced understanding of genre and the use of language, which defines genre as "typified rhetorical actions based in recurrent situations". Miller may have been one of the first to emphasize genre as something distinct from form and to argue for an understanding of genre as not being susceptible to taxonomizing because of its changing nature. Her ideas have been adopted and further elaborated by e.g. Swales (1990), Bazerman (1994), Mayes (2003), where Swales (1990, p. 49) admits that classification is an important feature of genre studies, but that genre should be determined by means of similarities, rather than by means of definitions or assumptions of prototypes.

For Swales (1990, p. 58), genre is "a class of communicative events, the members of which share some set of communicative purposes". It is these purposes that make up the "rationale" of a particular genre that "shapes the schematic structure of the discourse and influences and constrains choice of content and style". Genre is by no means defined in terms of form or style alone (p. 52). Genre theory is very explicit on the societal aspect of language and where the North American school of sociolinguistics[9] retains a strong interest in how the lexical and grammatical resources of language are used, genre theory shifts its focus towards the societal configurations of language use. Although genre theory is a very promising theoretical framework, our task of algorithmic genre classification demands a use of more fine-grained closed investigations of the observable features of texts, which these works on genre theory lack.

However, the genre theorists' interest in text structures is valuable. One of the most well-known examples of conventionalized structures of documents is the IMRD (Introduction, Method, Results and Discussion) structure of scientific writing, which has been the object of study in many cases, even outside genre theory. A good summary and

[9] See page 46 for an account of this school.

analysis of this is given by Swales (1990, Chap. 7). For another group of texts, Bazerman (1994) studies the structure of patent applications and grants, and investigates how the constituents of applications and grants may be seen as different speech acts. Here, we find indications of how something that is usually ignored by pure linguists attains interest: the document as a structure of constituents above the clause and sentence levels.

Swanson (2003, p. 21f) argues for a distinction where genre is associated with cultural context and the word *register* with situational context. The register, for Swanson, is a mediator of the realization of a genre. With the notion of register we come closer to the text. This notion has been extensively used by the Australian systemic-functional school, among others.

2.2.3 The systemic-functional view on genres and registers

In the systemic-functional grammar, as proposed by Halliday & Matthiessen (2004), the context plays a central role in the study of texts. It is in essence a social approach to the study of language, where words "get their meaning from activities in which they are embedded" (Halliday & Hasan, 1989, p. 5). Functional grammar may be described as a sociosemiotic grammar and is thus focused on a level closer to words than genre theory. The quotation above is surprisingly reminiscent of Wittgenstein's theory of language, where words get their meanings through their use in language. In fact, sociolinguistics as a whole, owes much to philosophies of language that stress the significance of context, e.g. to the works of Austin (1975) and Hymes (1974) mentioned above. Functional grammar is a formalism opposite to e.g. the Chomskyan context free grammars, where context and, therefore, meaning is left out in the grammatical analysis.[10] In fact, in place of grammar, Halliday prefers the term *lexicogrammar*, with which he explicitly marks the task of separating syntax and semantics as a disputable one.

The idea governing the Hallidayan systemic functional school is that situational and cultural contexts make up and maintain the struc-

[10] See Jurafsky & Martin (2000, p. 326f) for a description of context free grammars.

tural and systemic nature of human language. Different contextual settings result in a "register repertoire" of language, a set of functional varieties for language use, which are sometimes referred to as "genres" having "generic structure potential" (e.g. Halliday & Hasan, 1989, p. 64) . When language is expressed in texts of different types, the types of text are correlates to registers. Registers are varieties seen from the perspective of language potentials, the system; text types are varieties seen from the perspective of texts (Halliday & Matthiessen, 2004, p. 27). In Halliday's view, register and *text type* denote two concepts with different analytical perspectives on the same phenomenon.

Halliday's understanding of register has been criticized for being deterministic and (as an analytical instrument) suffering from the open-ended character of the variables used (Biber, 1994, pp. 33-34). For both Halliday and Biber the term "register" often refers to the functional varieties of a language, but for Biber the word is also used to denote something more general. Biber (1994, p. 33) uses it as "a general cover term for all language varieties associated with different situations and purposes".

However, in the case of both genre and register studies the question at hand is the use of language. Genre and register studies always relate to the intermediate level between language as socio-semiotic resources and its infinite variability in situated use. The task for genre and register studies is to make generalizations from language use and identify typifications of different kinds that arise from common contextual configurations — configurations in terms of e.g. culture, situation and purpose. In linguistics, the study of registers is mainly descriptive and the linguistic knowledge that derives from it may be used for educational purposes. This is in fact the explicit purpose of e.g. (Halliday & Hasan, 1989) and (Swales, 1990). Our purpose is, of course, to use it as a means of supporting information access.

2.2.4 Text typologies

From the perspective of information seeking, knowing that a certain document is about e.g. German grammar is generally not enough to determine its usefulness, or even its meaning. Halliday & Hasan

(1989, p. 45) make a very strong point of this, implying that there is no way of grasping the meaning of a text without its context. If a document is a thesis or a high school textbook for secondary language learners makes a big difference. Knowing that a certain document is about cars does not imply knowing whether it is an instructive text on how to repair particular cars or a text that describes this year's new models. The difference between a thesis and a textbook can be defined as that between a predominantly argumentative and a predominantly instructive text. Such designations refer to a fairly old system of categorization within the study of language that is referred to as text typologies. Ledin (1999, p. 18ff) gives a good overview of different approaches.

The term *text type* has been used in several contexts and its use is far from consistent. What follows here is an account of its differing uses in approaches that explicitly deal with text typologies. To begin with, let us consider the notion of what has been termed *text grammar*.

A certain text grammar, in the way Werlich (1976) understands it, can be characterized as a grammar that deals with the composition of linguistic units into coherent and completed texts. A text grammar thus extends the common linguistic occupation with morphemes, words, and clauses to include larger constituents and the composition of completed texts as sequential structures realized by different kinds of linking. Text grammars are also typologies that to some extent transcend the surface level of a text and take account for its inherent semantics (Ledin, 1999, p. 20). Thus, they are not independent of topic or theme. On the contrary, topicality is sometimes in focus, evene though the aim is not to identify the topics, but to investigate how the topics, among other text characteristics, contribute to the development of a text. From the perspective of Nunberg (1999), a text grammar is not only characterized by an extended focus towards completed texts, but to some extent different from a "lexical grammar" in that, for instance, a sentence of the text grammar does not necessary coincide with a sentence of the lexical grammar.

```
With a little help from my friend.
```

Example 2.2.3

This example contains a text-grammatical sentence, but as it is not a well-formed clause, being just a prepositional phrase, it is not a sentence from the perspective of a lexical grammar.

The systematic and extensive character of Werlich's text grammar focuses on text constituents that range from words to longer sequences of text. However, it should be kept in mind that the definition of text presented by Werlich (1976, p. 23) is rather restricted. A text is required to be a structure "marked by both *coherence* among the elements and *completion*" [emphasis in orig.]. It follows then, that badly written texts, unfinished texts and text fragments are somewhat out of the scope for Werlich's grammar. Moreover, Werlich's focus is not really the use of texts or what texts do, but on the linguistic competence required to produce a text. It is the relationship between the author and the text that is of interest for Werlich, where cognition seems to have a prominent place.

First of all, Werlich counts adherence to two exhaustive *text groups* as the primary distinguishing character. Texts are either fictional or non-fictional, which is determined by how the author relates the text to the context (extratextual fields of reference), i.e. "public time, locations in space, individual persons, fields of subject-matter etc" (p. 42). Fiction is "situationally autonomous" whereas non-fiction is not. Non-fiction presupposes that the reader ("the decoder") shares the author's ("the encoder") "referential presuppositions", fiction does not. Text types, on the other hand, are based on "dominant contextual foci" and Werlich distinguishes between five general types. A text type is either (in Werlich's own terminology)

1. a description (concerned with factual phenomena in the spatial context), that is often phenomenon-registering;

2. a narration (concerned with factual and/or conceptual phenomena in the temporal context), that is often action-recording;

3. an exposition (concerned with the decomposition into constituents elements, or the composition from constituent elements, of concepts of phenomena that the communicants have), that is often explicatory, identifying and linking phenomena;

4. an argumentation (concerned with relations between concepts of phenomena that the communicants have), that is often contrastive, attributing qualities to phenomena; or

5. an instruction (concerned with the composition of observable future behaviour, with reference to phenomena, in one of the communicants), that is often enumerative and action-demanding.

The attributes of these text types relate to coherent and completed texts and are abstract analytical categories, but they may also apply to smaller text units. Thus, a descriptive text unit may be linked to an argumentative text unit, while the complete text is characterized by its dominant type, partly determined by the character of its thematic expansion that can be either topical or functional, i.e. established on a semantic level (e.g.: 'The boat was approaching. The water was calm.'), or by means of function words (e.g.:'The boat was approaching. It was crowded').

Werlich (1976, p. 46) adds the notion of *text forms* to the text types, which are "conventional manifestations" of text types, such as "narrative, story, novel, report, or short story". Text forms in Werlich's text grammar are reminiscent of how non-fictional genre is defined by others, such as from the point of view of Grepstad, described on page 37 in this work. However, Werlich is not concerned with social and situational context.

Besides text groups, text types and text forms at the level of text constituents, Werlich uses the analytical categories of *point of view*, *composition* and *variety*, which he considers to be determinants for the completed text. For instance, a descriptive text form can be characterized as an impressionistic description if authored from a subjective point of view (indicated by sentences that express e.g. feeling). Such a text form may be composed of a descriptive introduction of a special "direction-determining" character to give the reader a good spatial impression of the phenomena described and their surroundings. Varieties are a category of a mixture of different qualifiers, such as idiolect, dialect and sociolect. The varieties also include style, which for Werlich is concerned with the author's responses to the phenomena referenced in the text, such as in an ironical or polite style. Werlich also mentions register but does not seem to consider register that much. He defines it as a "social role variety" tied to particular situations, and it may be due to Werlich's reluctance to take non-linguistic determinants into account that he does not elaborate more on register.

Werlich's typology is only one example of typologies that aim at categorizing texts according to their intrinsic properties. Robert Longacre and Jean-Michel Adam are examples of other scholars who have designed typologies of similar character. Although it is valuable to distinguish between instructive texts and argumentative texts, text types constitute fairly broad categories that do not necessarily have a direct correlation with genres.

The term *text type* is used in a different way by the historical linguist Manfred Görlach. Görlach (2004) has created a taxonomy of approximately 2000 named text types presented in an historical survey of the development of text types. Obviously, this notion of text types has little in commmon with Werlich's and is more comparable to the way in which documents are looked at by those working with markup languages (see Section 2.3).

A partial reason for Görlach's overwhelming number of text types is his notions of bound and free text types. Among the text types recognized are text types like 'captions', 'dedications' and 'prefaces', which usually do not occur independently of other text types — they are *bound text types*. Bound text types do not have to meet the requirements of coherence and completion presented by Werlich.

Görlach only counts *named* text types, which is rather debatable since text types (or genres) need not be recognized or named (Ferguson, 1994, p. 22). It is likely that the number of text types could in fact be much higher. Even though Görlach includes web pages, he does not mention FAQs or blogs. However, his consideration of bound text types demonstrates a necessary refutation of the monolithic approach that assumes documents to be unitary entities, especially with regards to web documents. Where, for instance, Werlich would probably regard a footnote as a non-text, Görlach would treat it as an instance of one bound text type worthy of longitudinal study — a task which is in fact pursued by Anthony Grafton in his *The Footnote: a Curious History* (1997).

2.2.5 Register studies

In the North American tradition of "register studies" Ferguson (1994, p. 21) discusses genre in terms of a "message type that recurs regularly", and Biber (1994, p. 52) refers to his own use of the word as a generic term for "text categorizations made by the basis of external criteria relating to author/speaker criteria".

One monumental work of register studies is the *Longman Grammar of Spoken and Written English* (Biber et al., 1999), in which is given a description of the actual use of English grammar by means of extensive corpus studies. This work penetrates grammatical patterns in terms of their functional aspects, which are of three main types: 1) their illocutionary force, 2) the constraints of language use that they express, and 3) the social and situational context they witness (Biber et al., 1999, pp. 41ff).

Register studies is further exemplified in a study by Kim & Biber (1994) on register variation in Korean. In a collection of 150 spoken and written texts, the frequencies of 58 different linguistic features are analyzed. The 58 features are reduced to a set of 6 dimensions by means of a factor analysis, that is, clusters of features that co-occur the most frequently.[11] For each dimension, then, the 22 predefined registers can be positioned on a continuum ranging from those that demonstrate the highest characteristics to those with the lowest characteristics of one dimension.

For instance, for the dimension of "fragmented structure" versus

[11] This is the same technique that Biber (1988) applied in his much cited work *Variations Across Speech and Writing* on register variation in English. The observant reader may have guessed that what is referred to here as "dimensions" roughly corresponds to the "facets" of Kwasnik and Crowston presented above.

"elaborated structure" we find, not surprisingly, (spoken) private conversations on one end and literary criticism on the other. This dimension corresponds to high frequency values of such features as contractions, demonstrative pronouns, relative clauses, lengthy sentences, and attributive adjectives.

The criticism raised against these studies, mainly coming from the point of view of genre theorists (Swales, 1990) and those working with genre identification on the web (Santini, 2004a), is that they apply a fairly restricted approach to the definition of genre, register or text type, where the space of typifications is taken to be *a priori* determined based on the view of the expert analyzer. Here, in the "new genre theory", genre is understood as constantly changing and being shaped by as well as shaping a discourse community. However, studies of this kind are extremely valuable resources for genre identification, since they provide us with clues as to which features may be the most discriminative for automated identification of genre.

Text typification is also relevant in the development of large corpora for linguistic research. In the categorical framework of the Stockholm-Umeå Corpus (SUC), the texts are categorized in terms of 1) "genres" as a label for the main categories, and 2) "domains" as a label for sub-categories.[12] Press reviews are here considered a genre further divided into the domains of books, films, etc., and imaginative prose a genre with domains such as general fiction and humor. Genres, then, seem to denote categories at a more general level than domains (cf. Wastholm et al., 2005). For the Lancaster-Oslo-Bergen corpus, "domains" are substituted by the term "sub-genres", where e.g. academic prose is further divided according to "natural sciences", "medicine" etc (Biber, 1988, p. 69). The genre-theoretical foundation for these categorizations is difficult to grasp.

2.2.6 Concluding remarks on genre in linguistics

In linguistics as a whole, genre is only scarcely recognized as an important concept, possibly due partly to the fact that, except for in a few

[12] A corpus may be distinguished from a simple collection of documents as being annotated and carefully sampled (McEnery & Wilson, 2001, Chap. 2).

subdisciplines, it is not text as larger units that are of interest. Genre in linguistics is, however, closely related to such concepts as text type and register, which in certain subdisciplines, such as text and corpus linguistics, as well as sociolinguistics, forms the object of empirical studies in which the aims are to correlate linguistic patterns to text typologies of different kinds.

Compared to LIS, there is much more to gain from linguistics for any precise conceptualization of genre and its application to classification, at least if we accept that genre variation correlates to a certain extent with text type or register variation.

Linguistics offer extremely valuable tools for the study of the relationship between language use and contextual issues. However, linguistics do not take much account of the influence of technology, something which is studied within LIS by e.g. Francke (2008), Crowston & Williams (2000), and which relates to what is to be discussed in the next section.

2.3 Technological perspectives: document types — document structures and markup technologies

In the previous section, documents have been described from a linguistic perspective as typified compositions of text units of a constative and performative character. In that respect, it is possible to distinguish a linguistic structure of relationships between units that characterize the artefacts of a genre, among which we can distinguish e.g. the IMRD structure of some kinds of research papers. Halliday & Hasan (1989, chap. 4) make an elaborated analysis of how structure is realized in certain speech acts. Therefore, certain linguistic constructs would allow for an approximate identification of a genre. For instance, there is a set of verbs that marks the occurrence of performatives of a certain type of speech acts, and there are certain verb tenses and pronominalizations that are more or less frequently distributed within certain parts of a text as well as within certain genres.

This structure of a text is to be found on a linguistic level of abstraction, but the composition of a document as an artefact relies in addition on extra-linguistic (or para-linguistic) and technological means. From a linguistic perspective, this distinction is referred to by Nunberg (1999) as that between "lexical grammar" and "text-grammar".

Besides the punctuation that separates sentences and clauses from each other, the compositional act includes other kinds of *markup* as well. Nunberg (1999, p. 17) refers to this kind of markup as "text-category indicators". While it would be possible to identify the separation between paragraphs from their linguistic expressions alone[13], certain visual means assists this part of text interpretation on a superficial level. Apart from punctuation, whitespace stands out as one of the most important and fundamental means whereby the identification of compositional units is accomplished.

In the same sense that a question mark has semantic value, namely to mark the preceding clause as a question, whitespace has its own semantic values, which are more or less easy to interpret for the human eye. The difficulties are due to the fact that whitespaces, as well as other kinds of markup, are ambiguous and that different whitespace expressions are used to mark the same compositional functions. For instance, it is as common to mark paragraph separation with a few space characters at the beginning of a paragraph as it is to mark it with a vertical whitespace between paragraphs. The indentation of a piece of text may mark the occurrence of a quotation, but it may also be used to mark something as not being part of the line of discourse in the text, such as an illustrative example of linguistic expressions, which is how it has been used in this text on e.g. page 33 (where a change in typeface has been added — another way to mark a compositional unit by typographic means).

Thus, previous sections of this study have emphasized that certain linguistic markers on lexicogrammatical and semantic levels may assist in the *structuration* of a text. This is also the case with what may be termed *paralinguistic markers*. The visual means by which the latter ones are distinguished is unfortunately highly ambiguous.

[13] The medieval principle of no spaces between words and no distinction between uppercase and lowercase letters, referred to as scriptio continua, is one example of where interpretative requirements are high.

However, visualization of a text must be technologically realized. A text needs to be enacted on some kind of media with the assistance of writing tools, such as pencils or computers. The latter technology has inspired the elaboration of markup theory.

2.3.1 Markup

Markup theory is intimately related to what has been termed markup languages, that are used for encoding electronic documents and which assists in the construal of document structure. Markup as a general term denotes several kinds of (paralinguistic) labelling systems for the encoding of texts, and has been the object for typological analysis. In an early influential article, Coombs et al. (1987) make a distinction between descriptive markup and other kinds of markup, such as punctuational and procedural. The descriptive markup is characterized as the most beneficient,as markup is then realized as labels that surround a certain text unit with mnemonic markers that indicate the (functional) semantics of that unit. The visual rendering of such units is a choice that is treated as secondary and left for later specification in a style sheet. The visual markup, sometimes referred to as the physical markup, is thus a derivative of the descriptive markup and the style sheet application.

Descriptive markup languages then, as being artificial languages, have their own syntax, defined in document grammars that are termed document type definitions (DTDs).[14] A DTD construes a formal and prescriptive grammar for documents of a certain type that shares a generalized and hierarchical structure of text units (see figure 2.1 for an example of a DTD and figure 2.2 for a snippet of text and markup). Natural language analysis is often taken as neing analogous to the DTDs of markup languages (Raymond, 1992, Sasaki & Pönninghaus, 2003, McKelvie et al., 1998). The document grammar corresponds to the set of grammatical rules that defines well-formed clauses in a particular natural language. The generalized grammatical structure corresponds to the empirical or intuitional knowledge of what con-

[14] There are other formalizations than DTDs, such as XML schemas, but that is not our concern here, they have similar functions.

stitutes a well formed clause. Subsequently, the document structure corresponds to an instance of a well-formed document. speaking, formally there is no markup language until such grammars have been specified.[15] A poem, in e.g. the TEI framework[16], is expected to conform with its specified grammar in order to be counted as a poem. The DTDs, the encoded documents together with the style sheet are sometimes said to constitute the document architecture for "creating and processing a class of documents" (Lubell, 2001). As the quote indicates, a class of documents in this case is similar to the notion of a class of genre artefacts. It is only a much more coarse-grained notion.

```
<!ELEMENT Chapter (Heading, Paragraph+,
  ListOfReferences)>
<!ATTLIST Chapter id ID>
<!ELEMENT Heading (#PCDATA)>
<!ELEMENT Paragraph (#PCDATA)>
<!ELEMENT ListOfReferences (#PCDATA)>
```

Figure 2.1: Document grammar

```
<Chapter id='C1'>
  <Heading> [Some text]
  </Heading>
  <Paragraph> [Some text]
  </Paragraph>
  <Paragraph> [Some text]
  </Paragraph>
  <ListOfReferences> [Some text]
  </ListOfReferences>
</Chapter>
```

Figure 2.2: Document instance

The hierarchical nature of the structure imposed on any text by most

[15] It has to be emphasized though, that XML allows the XML syntax to be used without specifying a grammar.

[16] The TEI is, simply stated, a set of DTDs for the transcription and encoding of primarily cultural heritage material.

markup languages has been much criticized but remains a significant feature of most markup languages. This is contrary to, for instance, the superficial IMRD structure referred to above, which may be seen as a flat sequence. Since markup languages concern structures, and markups may be of different kinds, it is possible to distinguish between what Peels et al. (1985, pp. 347-348) term a logical and a physical structure, where the former refers to the semantic structure and the latter to the structure rendered on some medium. The logical structure is definitely a compositional hierarchy (McKelvie et al., 1998, p. 368), but whether the same thing could be said about the physical structure remains a question. (A sample of a logical document structure is given as a dendrogram in figure 2.3.)

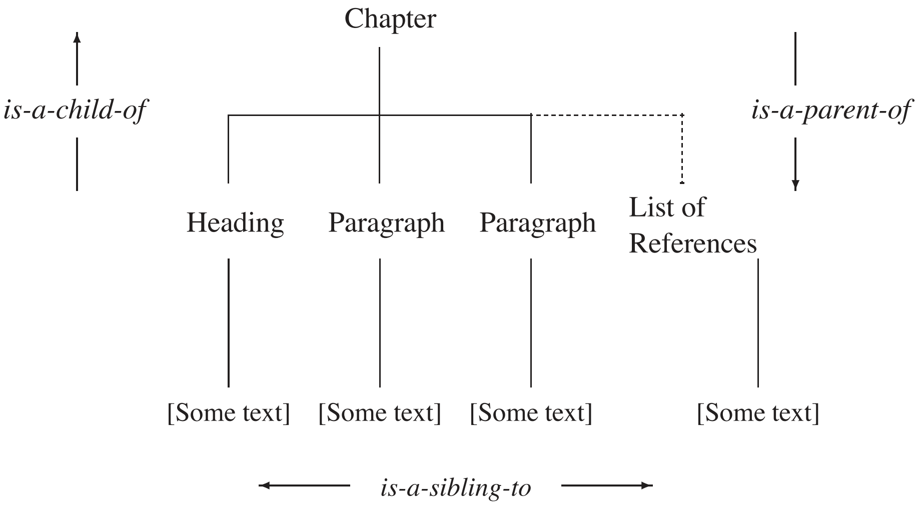

Figure 2.3: Compositional hierarchy
Note that text nodes are not in sibling relationships with each other.

However, at present there is no standardized way in which the semantics of markup may be specified, beyond what may be inferred from the names of the labels. Such attempts have been made recently by Dubin et al. (2003), and Renear et al. (2002). Renear (2001) has also proposed a dismissal of the distinction between descriptive and procedural markup and attempted to substitute it with a theory based on artefactual text units as expressing different kinds of speech acts. For instance, the encoding of a title in case of a transcription procedure, where the task at hand is to describe the source text, could be considered an indicative (or constative) act (i.e. it is either true or false with respect to the intention of the author or publisher of the source text). The encoding of a title in a source text, on the other hand, could be considered a performative, because it enacts the intention of making a title. It is neither true nor false. The encoding of bold text, finally, could be considered an imperative act, in the "renditional" domain, as opposed to the "logical" domain to which the encoding of a title always belongs. Thus, just as linguistic utterances may be considered as speech acts, markup may also be considered a speech act, and not only descriptive. What is to be remembered from this is that markup is a far more complex matter than it is sometimes depicted and reflects text production as action, or as Sperberg-McQueen (1991) put it, markup "reflects a theory of text".

Markup of different kinds always assists in defining text units. If it relies on an underlying hierarchical model, it defines a text as a hierarchical structure that is not independent of the linguistic text structure that is being encoded. However, the relationship between the two structures is far from clear-cut. It may be expected,however, that where there are nodes defined by markup, there is also some kind of semantic cut between content units. A fact that makes it possible to identify units in algorithmic ways from the mere markup.

Among the inherently descriptive markup languages there is one that has met an astonishing success, namely the HyperText Markup Language (HTML), which is one of the technological foundations on which the web rests. The majority of documents distributed on the web are encoded in some version of HTML. The heterogeneity of the HTML part of the web, with respect to language use and genre, is far more apparent than for any physical library. Unfortunately, even though HTML is defined as a document type, its coarseness does not lend itself to direct genre classification. Even so, the use of HTML opens up an opportunity for algorithmic application, since HTML relies on the combination of ASCII character representation and markup.

As there is nothing mandatory in working with HTML encoding isolated from the specification of style sheets, and a default style sheet is applied if none is provided by the author, HTML is mostly considered presentational, a formatting language. The benefits of a descriptive markup language, being fairly unambiguous, have been mostly ignored by the communities that use it. While it could have been expected that e.g. an `h1`-tag always marks the occurrence of a heading, and that such occurrences would have been used to model the logical structure of the document, this is far from certain. With respect to the implicit semantics of HTML grammar, *markup performance* is in general deficient. This is partly due to the inherent generality of HTML. HTML was defined to be used for almost any kind of document, any thinkable genre (Berners-Lee, 1989).

As has been touched upon already, it is fair to assume that even though the labelling of a text unit cannot be trusted, the unit defined by enclosing tags is a text unit of some kind, or a part of a text unit, even with respect to a logical document structure. It is just that the semantics of a text unit is implicit, if it has a meaning at all.[17] There is ambiguity in the markup performance with respect to document structure semantics, but this does not mean that we necessarily have a constituent ambiguity as well.

2.3.2 Markup theory

Markup is one possible source for feature derivation in document genre classification and it is therefore appropriate to clarify some of the concepts and terms used in markup theory.

Figure 2.3 depicts the (abstract) document structure of a snippet

[17] Since the writing tools (i.e. the HTML editors, such as MS FrontPage) often hide the encoding, meaningless markup may accidentally occur.

from a possible text of some kind — a chapter containing a heading (or a caption), several instances of paragraphs and a list of references. The latter constituents are contained in the chapter, thus being in a child-parent relation, and the chapter in a parent-child relation to its children. Each of the children to the chapter is a sibling in relation to the other children. This genealogical metaphor is common within the discourse of XML application and it should also be noted that the tree structure may be instantiated to any depth, thus demonstrating fairly distant ancestry relationships.

For the snippet presented in Figure 2.3 we would expect a grammar that defines a generalized content model for the chapter and its children that may be realized as in Figure 2.1. Line 1 declares that a chapter must contain a heading followed by one or more paragraphs and end with a list of references. Line 3 defines that a heading contains text and nothing else.

Looked upon as an abstract document structure, the individual components of an instantiation are referred to as *nodes*. The nodes are instances of the *elements* defined in the grammar. The word node is sometimes used synonymously with the word element.

It is worth noting that we usually picture headings and paragraphs as containing text, but the elements of this abstract construct sometimes do not contain text as direct descendants — as children. This may be true in case of composite text types, such as textbooks, which may contain nothing but other element instances. It is also true of the chapter element in figure 2.3. Markup theory makes an important distinction between these types of nodes, and refers to them as *text nodes* and *element nodes*. A text node is just a piece of text and an element node is an abstract container that may or may not contain text nodes. There are more types of nodes, but it suffices to mention attribute nodes, created by the use of an attribute. In figure 2.1 the attribute is merely intended to be used in giving any occurrence of a Chapter-element a unique identifier.

A document is realized by markup that consists of *tags*, constitued by an element name in angle brackets. This name functions as a *generic identifier* and refers back to the element name defined in the formal grammar. The particular grammar given in figure 2.1, in combination with the abstract document structure given in figure 2.3 and the conventional nature of XML markup notation, would yield an instance snippet as exemplified in figure 2.2.[18]

2.3.3 Three structures

As has been touched upon already, a document structure may be identified on different levels. First, there is a visual level of a physical (or graphical) document structure perceived by the human eye. Second, there is the level of the logical document structure that concerns the composition of functional text units into meaningful artefacts, conveyed by markup. Third, there is the linguistic or rhetorical document structure, predominantly expressed through a process of textual interpretation in which the two other structures assist. The second level, which is inherently technological, has very often been ignored in linguistics. An interesting exception is Power et al. (2003), who argue for the implementation of the second level in document generation engineering.

We may expect genre typification on all these three levels. Typification on the linguistic level(s) has been treated in Section 2.2. Typification on the visual level is a more uncommon theme. Ihlström & Åkesson (2004) have studied Swedish online newspapers and analyzed their front pages from the perspective of a tabular grid. They found them fairly consistent with their printed counterparts. The role played by the intermediate document structure level, conveyed in markup enactments, is however fairly uninvestigated.

Decomposing an HTML document based on markup results in a multitude of nodes. Among the nodes, the text nodes are of greatest interest, as they contain the textual contents. This decomposition results in a very fine-grained structure, somewhat midway between a clause or sentence level and a level of coarse structural elements, such as introductions and results sections in a research paper.

These levels of granularity are illustrated in Figure 2.4, where the horizontal lines represent a document looked upon as a sequence of

[18] Some mandatory notational curiosities have been omitted in order to avoid unnecessary confusion for the reader.

broad functional text units (corresponding to Görlach's bound text types), text nodes, and linguistic units respectively, and the vertical lines show the different levels of decomposition.

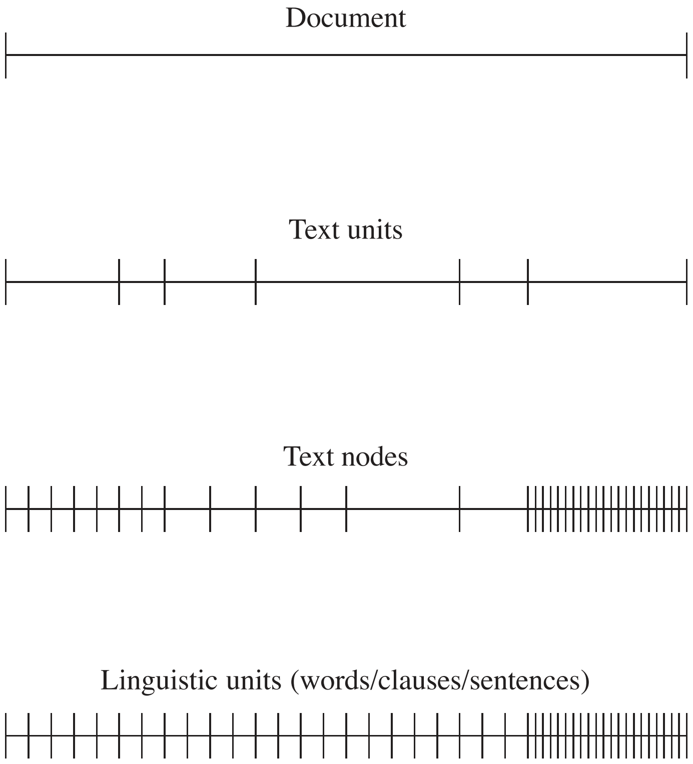

Figure 2.4: Levels of decomposition

2.3.4 Concluding remarks on document types

The notion of document type in the context of markup theory must be considered a facility for defining a technological counterpart to classes of genre artefacts, however most often within a far more coarse-grained genre space. The definitions prescribe a structure above or below the linguistic structure (depending on our perspective) that fits the requirements of certain documentary practices. Thus, given a certain set of DTDs and a set of documents fully conformant with the DTDs, and any recommendations on how to apply them, this fact would be sufficient for a coarse-grained classification based on mapping the DTD declarations to classes. A document that conforms with the Verse or Drama TEI subsets would be easily identified and classified from the markup only, as would a document conformant with the Dictionary TEI subset. A more striking example is probably the DTD promoted as a formal American Standard for *Scientific and Technical Reports* by ANSI/NISO Z39.18-2005, that is intended to "...foster uniformity ...for ease of information retrieval ..." (American National Standards Institute, 2005, p. VII).

Unfortunately, the situation is that the majority of documents in need of organisation are encoded in some version of HTML, without being neither valid nor well-formed. The HTML encoding itself would need disambiguation in order to be used for classificatory purposes. Still, the compositional structure resulting from the HTML markup cannot be ignored as to its potential for document feature derivation.

2.4 Towards a theory of genre

It is not possible to formulate a set of theoretical statements on genre that fully comply with all the ways in which genre has been treated in this chapter. Anyhow, this section will attempt to state a few things that only to some extent refute what has been said on the notion of genres in the theories so far recapitulated.

From the perspective of LIS, documents can be considered as intermediary artefacts in documentary practices. They are inherently social objects intended to do some work in human interaction and may be used by humans for many different tasks. Since LIS is concerned with appropriate access to documents within larger collections, it is assumed that bibliographic practices (e.g. classification) must not be solely confined to the topical characteristics of documents, but also to their non-topical characteristics, particularly related to the use of language and technology. The notion of genre as social action provides an appropriate framework for these objectives.

From the perspective of linguistics, documents are seen as compounds of both constative and performative linguistic utterances, the latter being the constituents of text units.[19] Text is produced in order to say something (a locutionary act), but also to accomplish a task (an illocutionary and perlocutionary act). Even though the use of natural language may theoretically be said to be infinitely variable, its patterns become highly typified by contextual factors related to the varying situational, technological, and sociocultural settings, whose typifications are the focus for this work — expressed in terms of genres.

From a technological perspective, documents are derivates from the use of technology and any attempt to characterize documents must account for restrictions put on tehir creation and use, and the possibilities offered by technological innovations. Technological innovations, especially those related to the growth of the web, are most often seen as the causes for "the formation of novel genres, genre hybridism, individualisation, and intragenre and inter-genre variation" (Santini, 2007, p. ii).

The term genre indicates an area of study where LIS, linguistics and technology studies may converge. A document relates to genre at that abstract point where the text and the artefact connect to the situational and sociocultural conditions around its production and use. Documents are not primarily transparent objects that convey human knowledge, they are objects inscribed by human action and thus imbued with traces of the more or less conscious choices made in a certain situation within a certain community of practices. The full meaning of a document cannot be inferred merely from its references to abstract or material phenomena and events independent of the document. The purpose of, for instance, a contract between two parties is not only to convey knowledge but to regulate future interactions between the parties involved, enacted by the mere existence of the contract. It is an object of social negotiation and mediation. While there is almost always a topic that motivates a document's existence, you can-

[19] This work considers text documents only, even though the methods and theories may be slightly modified to reflect an increased generality, necessary for accounting of non-texts.

not assume that its appropriate use is dependent on what a document is intended to do. The distinction between, for instance, on the one hand a predominantly argumentative and on the other hand an instructive approach behind the production of a document reflects variation in the communicative purpose. And, as has been pointed out, genre is the common denominator for communicative purposes.

Successful communication requires that the response to an argumentative act of communication must be different than that to a purely instructive act, and thus determine the usefulness of a document with respect to a user-defined task. Since documents are seen as a part of human interaction, we have a kind of action that involves artefacts as the means of communication — *a documentary act*.[20]

The communicative purposes answer different questions. If we need to know how to do things, an instructive text is more useful than a predominantly argumentative text, whereas an argumentative text is probably better for making well-founded decisions in complex situations. Scientific writing which, within the social sciences in particular, can be characterized as mainly argumentative, needs citations to back up the arguments put forth. Many of the citations refer to texts and authors that are summoned to ensure that the arguments are worth trusting — the author's allies, as the sociologist Latour (1986) puts it — and need to be carefully chosen. The texts chosen as being appropriate depend on the positions held by their authors and publishers within the community of discourse at hand, as well as on the character of the texts, e.g. their genre adherences.

Communicative success as a whole depends to a large degree on the author's and reader's ability to recognize genre conventions — to compose and interpret the documents against the background of typical genre constraints (cf. Vaughan & Dillon, 2006, Halliday & Hasan, 1989). There is no coincidence that much of the recent attention given to genre and its practical applicability has been within the area of language learning for special purposes (for instance, Swales, 1990, Halliday & Martin, 1993, Bazerman, 1989), as composing texts according

[20] The immediate connotations of this term may seem more restricted than intended. The term does not just refer to the production of a document, it also refers to its circumscribed sociocultural actions and technological conditions.

to reader expectations is an important communicative competence.

Genre should be understood as an abstract phenomenon, not as document form, if its importance for information access is to be respected. Form is important, but it only reflects genre. Assigning the correct name to a class of documents is not enough, unless what the form bears witness to is known. A description of a genre may very well be more wisely made if the situation is thoroughly described, rather than resting on ambiguous names of genres. This is actually what Halliday & Hasan (1989) do when they break down a piece of text according to an analysis of its field, tenor, and mode. The term genre denotes a conflation of several document-intrinsic and document-extrinsic configurations that relate to its social function and the documentary act it reifies. From this approach, a genre can be analysed on three analytical genre dimensions, where different genres that are recognized and possibly given names do not always lend themselves to being fixed with respect to each of these dimensions.

- The community of discourse and practices circumscribing a document.

- The communicative situation to which the enactment of documentation is a response, where the act is performed with more or less conscious purposes

- Artefactual typifications that reflect the documentary conceptualizations and expectations related to particular communities, situations and modes of communication

This may be graphically simplified as a combination of triples (Subject-Predicate-Object) in Figure 2.5. For instance, how a situation is configured *motivates* a certain genre, which in turn *generates* documents, and the expectations *form* the genre.[21]

An important reservation to be made here is that even though the unidirectional arrows seem to imply that there is a unidirectional influence from e.g. situation to genre, this is only coincidential. A genre,

[21] The resemblance with the directed labeled graphs of the RDF model is obvious, but should not be strictly interpreted in that way. For instance, the ovals do not represent addressable resources as they do in RDF.

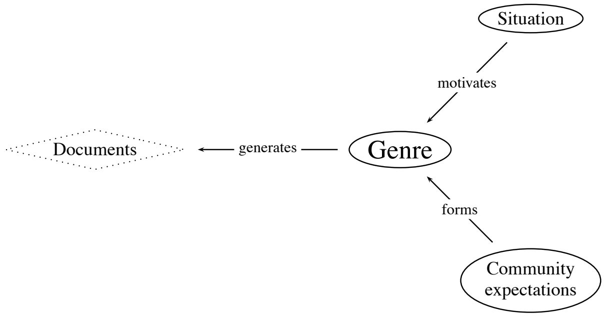

Figure 2.5: The genre triples

or rather the documents, influence the situation, by means of, for instance, intertextual relationships. It is true that a situation motivates the initiation of a genre, but an existing genre also affects the situation by placing restrictions on it. For instance, the transfer of real estate ownership between two parties cannot proceed without the parties taking account of all the elements of the different documents needed for the transfer of ownership in their particular culture. The unidirectionality only reflects the labels that predicate the triples.

Proceeding with an example, the name `bibliography` is assigned to classes of documents that list references to works cited in a text, but also to classes of documents that are the result of what Dahlström (2006) terms material bibliography ("materialbibliografi"), where the purposes are completely different.

The case of the label `bibliography` points to the problematic issue of names that may be rather confusing when used as genre labels — if we admit that the name bibliography is a name of a genre, that is to say.[22] Just as the word "classification" may denote both an entity and an activity (Jacob, 2004, p. 522), the word "bibliography" may denote both the product of bibliography and the activity of bibli-

[22] There seems to be no reason not to. The "web genre community" which has put up a Wiki (WebGenreWiki, 2008) for the discussion of classification along genre dimensions issues, lists both bibliography and link lists among the recognized genre labels.

ographing. The outcome of the activity of bibliography may be what is adequately termed a bibliography (a listing of books), but it may also be, e.g. , a critical edition that makes up a kind of archaelogical research project. The fact is that the purposes of a bibliography in a scientific work, a national bibliography, a library catalogue, and the works of material bibliography are very different, even though they may all be said to be the outcome of bibliographic activity.

A national bibliography is the outcome of enumerative bibliography and provides primarily a documentation of what has been published in a certain language and/or country. It can, among other things, be used to determine what has been published, but it is not a finding aid in the same way as a library catalogue, and certainly not in the same way as a bibliography of e.g. a thesis. The latter one is produced in order to link a scientific work to allied researchers, and not primarily to direct a reader to further reading, as is the case with some bibliographies produced in libraries.

Thus we see that the name bibliography is usually given to a class of documents that are listings of other documents. Here, it seems that the label bibliography is given to documents that share no other common trait than having similar form. Anyhow, since they are similar with respect to form, they are often recognized as one genre, even though, according to the triple, they should not be.

A clearer evidence of the fact that the goals and intended readers of a document are more important than form is the PhD thesis. One of the most prominent features of theses is that they serve the purposes of gaining degree within a certain academic discipline. Their form, however, may vary a lot, depending on the academic domain or the topic. In a PhD thesis there is usually one feature that is almost always present, namely a verbal statement of the fact that it is a PhD thesis. The most important fact for its use, however, is that it relates to the academic degree.

A consequence of this way of defining a genre would in fact be that it is not strictly appropriate according to the kind of linguistic tests referred to in Section 2.1.1. A statement like `this book is a bibliography` is not a statement of its genre, but on its adherence to a class of documents usually referred to by that name. A common but sometimes confusing fact is that names given to classes of documents are sometimes in fact designators of something other than a class of documents. In the WebGenreWiki (WebGenreWiki, 2008), a wiki set up for collaborations and discussions between (web) genre researchers, a list of genres is suggested, where one encounters, e.g., `Public debate` and `Petition`, labels that really indicate types of actions. However, one also encounters labels such as `Homepage`, `Law`, and `Glossary` that have to be considered labels of classes of documents tied to activities of portrayal, legislation and definition, respectively. An odd item in this list is `Pornography`. Thus, names are often confusing in terms of an understanding of genre but are nevertheless deeply rooted in common language use. Hopefully, any analysis of documents performed on all the three dimensions would serve the purpose of clarifying relationships of importance for information access.

However, it is with respect to the last of the dimensions above that the patterns of the artefact reify the functional role that it has in a communicative and documentary situation. The mediating characteristics of this dimension may be found as typified patterns on different levels of abstraction and different levels of intrinsicality, ranging from e.g. choice of technology to influential external factors derived from the domain in which the artefact plays a role — such as rules of conduct specific for a user community. We would, at least today, be surprised to find a shopping list carried on an e-book reader, a shopping list written as a narrative, or a conference organised around a set of artfully designed shopping lists. Documents are tied to the triple of a specific communicative situation, a sociocultural configuration, and specific accepted forms of expression, but it is most probably only the last one that may be susceptible to algorithmic processing. Classification algorithms have to be built on the observable realizations of genres. The next section will summarize the ways in which genre may be approximately recognized.

2.5 Recognizing genre

It has repeatedly been emphasized that the notion of genre adopted for this work is an abstraction. However, the observable properties of a document are assumed to be typified with respect to the extrinsicality of its communicative function and discourse community. Thus, it should be possible to proceed from these observable properties to statements of genre adherence. The solution lies in the selection of observable features and an investigation of their covariance, for the purpose of which linguistic research on register and text types provides valuable background knowledge. This section aims at sketching out how a few research approaches depicted in the previous sections may be used.

First of all, when Austin's speech act theory is considered, it is inevitabel to note that verbs in general have certain performative meanings. For instance, it is obvious that verbs in the imperative mode are primarily directives aimed at having the addressee to react upon what is said. Questions are another kind of constructs of obvious performative nature. Both interrogatives and imperatives have been considered non-assertive, even before Austin. Austin's observation that a combination of the pronoun 'I' and a verb in the present tense, where the verb pertains to certain categories (commissives, verdictives etc), is an obvious example of a performative. Thus, verbs are particularly interesting as indicators of communicative purpose (cf. Austin, 1975, Chap. 12).

The construct "we define" may serve as a good example. 'Define' is, according to Austin, an expository verb.[23] In a search performed on Google with the phrasal query "we define", we get a set of document references where some documents' co-texts for the phrase are given in Figure 2.6. Because of the uncertainty of the weighting principles applied by Google it is not possible to draw any conclusions from this example, but it is interesting to note that among these 10 examples, 7 of them are clearly related to academic practices, within which domain expository and argumentative texts dominate.

[23] Thus it correlates to a certain extent with the expository text type.

```
1.  -0600 Previous message:  how would we define
“kin”?  Next message:  how would
2.  heap variables.  In this paper, we define a
parametricity semantics for a
3.  some new lines in .  First, we define some
constants and structures, somewhere
4.  We must be careful when we define matrices over
R to make
5.  gt; Corporate philosophy > How we define quality
Groz-Beckert Products Agencies Service
6.  peter193710 Level 1 How can we define
“intellectuals” and why are they
7.  DEMO Letter How should we define Web 2.0?  Chris
Shipley
8.  generate all vowel vowel diphone we define a
carrier (set!  vv-carrier ’((pau
9.  natural language extension:  just as we define
recursive functions on values using
10.  up the question of how we define the time
intervals we measure
```

Example 2.5.1

Figure 2.6: Co-texts for a ’we define’-query

Subsequently, as opposed to in the context of topic detection, lexical categories other than nouns are necessary to account for, as well as certain syntactical constructs. Apart from previous research on genre classification and identification, it is crucial to consider empirical research as expressed in register studies.

There is an important observation made by Biber (1988). Biber's work builds on factor analysis of linguistic features in the Lancaster-Oslo-Bergen (LOB) corpus, which is a collection of 500 samples from both spoken and written language use. He concludes that when considering coarse-grained genres, such as 'academic prose' or 'fiction', the variation within these genres is not coherent — at least not with respect to the underlying dimensions identified through the factor analysis. One has to consider more fine-grained genres (Biber, 1988, p. 170). This is, in fact, also what makes it appropriate to draw a sharp distinction between genres as extrinsically derived categories and text types as intrinsically derived categories.

However, Biber's investigation demonstrates some undisputable characteristics which also are consistent with common opinions. For instance, the frequency of temporal adverbials in "personal letters" and in particular in "broadcasts" are significantly high, whereas in "official documents", there seem to be no hedges (e.g. 'at about', 'more or less', and 'maybe'), and the frequency of adverbials in general is much lower than in e.g. "news reports" and "fiction" (cf. the tables in Biber, 1988, App. III).

It is supposed that a text type is the realization of genre at the artefactual level, but it should be kept in mind that certain genres may deliberately paraphrase a text type that does not normally realize the genre with which it is typically associated.

A striking example of an unusual text type used within a normally argumentative genre is the rewriting of a fairy tale by the Swedish author Astrid Lindgren in 1976. The Swedish newspaper Expressen, on the 10th of March, published her tale "Sagan om Pomperipossa i Monismanien" as a protest against the taxation legislation in Sweden at that time. Here, a narrative is unexpectedly used as an argumentation within a public debate on taxation principles connected to the future parliament elections that autumn.

However, this is an exceptional case and genre may very often be approximated from the use of language and technology.

Finally, to conclude this section on genre, in almost every approach to genre it is admitted that genre artefacts are partly defined in terms of their compositionality. Documents are compositions of constituents above the clause level, whose compositions may be equally typified as the use of language. For instance, for the broad category of "research articles", it was not until the 1930s that articles in general came to be extended with the higher-level constituents of sections for Discussion and Conclusion. Until then, Swales (1990, p. 115f) asserts, if sectioning was applied, articles usually ended with the Results section. In addition, headings were not common before that time. This is an indication of how genres change even with respect to compositional principles, however still being characterized by them.

Görlach's notion of bound text types is an important contribution to any attempt to study compositional units above the clause and below the document itself. Unfortunately, the matter of compositionality has not been given the same attention as the use of linguistic patterns within genres.

Chapter 3

Classification

In the preceding chapter it has been concluded that even though genre is far from a straightforward notion, there are many indications on that linguistic and extra-linguistic patterns correlate with different contextual and situational configurations in an apparently predictive way. The works of Biber and others within the North American school of register studies, reviewed in Section 2.2.5, probably serve as the strongest indications that this is the case for linguistic patterns. Therefore it also seems worth assuming that these patterns can be used in applications for the classification of documents along genre dimensions. However, how classification may be accomplished in general has not been investigated, which is among the things that will be considered in the following chapter.

This chapter focuses on the second research question that asks how the classification of documents along genre dimensions can be realized. This is accomplished by examining how classification can be formally defined and the different ways in which classification can be and has been realized and motivated.

Even though this is a work that positions itself within LIS, this chapter will start out from the algorithmic point of view on classification, rather than from the LIS point of view.

One reason for this departure can be given right from the start. As has been reported on in Section 2.1.3, there are within LIS a few attempts to establish classification schemes for genre, but these studies are not primarily interested in how the schemes may be applied, and certainly not with respect to algorithms. As for classification theory in general within LIS, most of the focus is on classification as a descriptive activity, rather than as a subdividing activity, and on classification schemes as stand-alone structures that can be studied without any particular collection of documents in mind.

This chapter will proceed from the simple statement of classification, given at the very beginning of this thesis, i.e. *dividing large collections of documents into groups of similar documents*. From this statement it will sketch out several ways of designing the task as a computational model.

There are several areas of research that deal with algorithms for classification in general and the classification of documents in particular. The domain of *machine learning* is an area of applied research that is sometimes counted as part of artificial intelligence. Machine learning relates to classificatory problems that do not necessarily involve texts or documents. An overview of machine learning at large is given by both Mitchell (1997) and Alpaydin (2004). Principles for machine learning in general are applied in many different contexts where classification is a prominent issue, such as in speech recognition. The domain of *text categorization* applies machine learning solutions to the processing of texts, such as junk mail filtering, authorship attribution or word sense disambiguation. A comprehensive introduction and overview to text categorization is given by Sebastiani (2002, 2005). *Information retrieval* is, as is pointed out by e.g. Sperberg-McQueen (2004), a special kind of on-the fly classification where a set of documents is divided into relevant and irrelevant documents with respect to a user query. Prominent introductions are given by Baeza-Yates & Ribeiro-Neto (1999) and Manning et al. (2008).

3.1 Defining the classification task

What is given in any classification task is that we have a set of objects X that are to be mapped to a set of classes C. In that respect, a classification might be considered a function Φ from X to C, so that $\Phi : X \times C \rightarrow \{0, 1\}$, where 1 refers to a positive assignment to a $c \in C$ and 0 a negative assignment. This means that for any $x \in X$ a classifier Φ has to decide, with respect to each $c \in C$, whether x is to be assigned to c or not. A seemingly odd way to redefine the task, which is sometimes done, is to not assume a binary class adherence and redefine it as $\Phi : X \times C \rightarrow [0, 1]$, where an object is assigned in degrees of adherence to each class. However, these degrees are generally resolved into binary values in the end (Sebastiani, 2005, p. 114), so they are mostly computationally and not theoretically motivated.

Therefore, document classification can be seen as a mapping of documents in a collection D to a set of classes C. Equation 3.1 expresses this.

$$\Phi : D \times C \rightarrow \{0, 1\} \tag{3.1}$$

This definition does not restrict the classification to a disjoint, or single-label, classification. It allows for an object to be included in several classes, what is sometimes termed multi-label classification (Santini, 2007, Sebastiani, 2002, p. 4). In order to restrict document classification to being disjoint it is best reformulated as in Equation 3.2.

$$\Phi : D \rightarrow C \tag{3.2}$$

In other words, document classification can be either *disjoint* (single-label) or *overlapping* (multi-label).[1] If we imagine documents as points in a two-dimensional space, and classes as parts of this space populated by a certain portion of the documents. Figures 3.1 and 3.2 illustrate the difference.

Furthermore, a classification task can be conceived of as either exhaustive or non-exhaustive. When it is exhaustive Equation 3.3 holds, and for a non-exhaustive classification it does not, insofar as at least one document is not assigned to any class $c \in C$ at all.

$$\begin{aligned} &D = \bigcup\{c_1, c_2, ..., c_n\}, \\ &c_i \subset \mathcal{D} \end{aligned} \tag{3.3}$$

[1] For consistency issues, the terms "disjoint" and "overlapping" are preferred in the following, because classification as a subdivision is considered more important than classification as a label assignment.

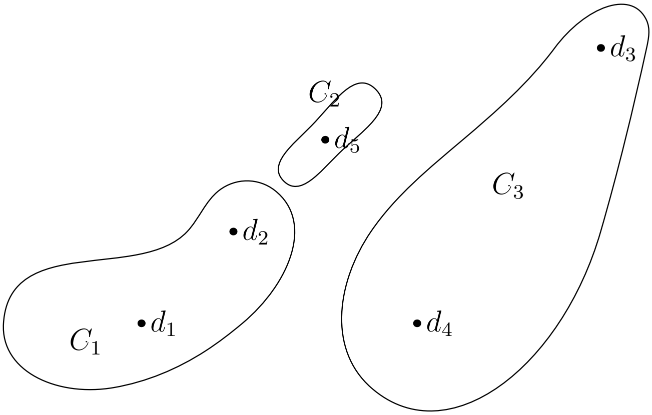

Figure 3.1: Disjoint classification

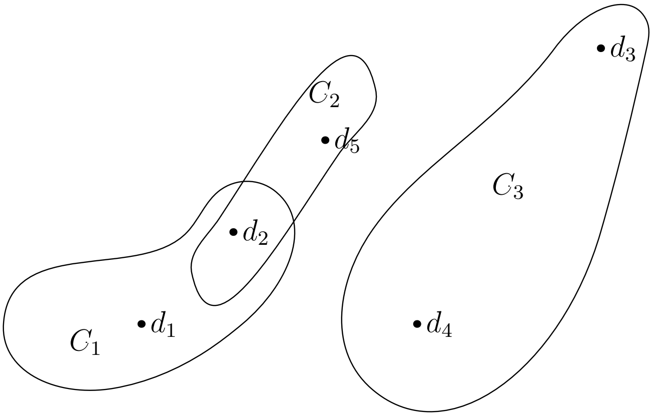

Figure 3.2: Overlapping classification

Having identified the distinction between disjoint and overlapping classification as well as that between exhaustive and non-exhaustive, the next issues concern the precise elaborations on both D and C. Definitions 3.1 – 3.3 imply a set C, but do not say anything about its character, or the character of each $c \in C$. Any ordinary algorithm for classification reported in the literature requires that at least the cardinality of C is defined *a priori*, i.e. one has to decide how many classes to which documents can be mapped.[2] In this work, classification should be understood as a task where at least the number of classes is predefined.

This restriction makes it possible to calculate a mathematically defined task complexity, based on the entropy measure. Entropy is calculated from the distribution of objects in X over the set of classes C. The entropy is then defined as $\mathcal{H} = -\sum_{c \in C} P(c) \log_2 P(c)$, where $P(c)$ is the probability for the assignment of an object $x \in X$ to class $c \in C$, the so-called *prior probability*. However, even though task complexity measured as entropy is bound to increase when the number of classes increases, it does not reveal much of the real task complexity, neither for algorithms nor for a human mind. When humans perform classification, the cardinality is huge but the actual task complexity is somewhat restricted in practice by the fact that classes, or rather, their representations, form a structured scheme with mnemonic notation and support for conceptual mapping (see Section 3.4.1 for an account of classification schemes and associated notation schemes).

In general, when algorithmic approaches are at hand, each $c \in C$ tends to be defined more or less by the documents assigned to it, as we shall aslo see later on when discussing different classification algorithms. However, according to the first general definition, classification relies on assumptions of similarities between documents, so that, typically, documents in the same class are supposed to be more similar to each other than to documents in any other class. This can be conceived of as a statement on what constitutes a document class. It also means that the definition of classes in algorithmic approaches is not independent of the ways in which decisions on document assignments are made. If the similarity between two documents is denoted

[2] There are a few possible exceptions to this, as some unsupervised classification models do not necessarily require this.

$sim(d_j, d_k)$, one may imagine a statement on a class c as in Equation 3.4.

$$c_i = \{d_1, d_2, ..., d_n | \forall d_j \in c_i, sim(d_j, (d_k \in c_i)) \geq sim(d_j, (d_k \notin c_i))\} \tag{3.4}$$

The definition states that for each document in a class it holds that it is more similar to any other document within that class, than to any document in another class. This measure will be referred to as the *harmonic quality* of the class formations. It will, however, become clear that it is not in itself a sufficient measure of classification quality, even though it is an important one, and also an ideal condition that must in some cases be violated.

It must also be emphasized that given this harmonic quality as a measure of quality does not imply that similarity measures must be the means for arriving at class assignments. For instance, the probabilistic Naïve Bayes classification model does not use similarity measures, but calculates probabilities for a document being assigned to a particular class, and assigns a document to the most likely target class.

What can be understood, though, is that $sim(\cdot)$ is a measure that has to be assigned a value, and the crucial question is how to arrive at such a measure. This is usually accomplished by some kind of decomposition process where one or more of the properties of a document are assigned values, usually termed attributes or features. Each document $d \in D$ can then be represented as a set $\mathbf{x}$ of features $x_1, x_2, ..., x_n$.

$$d \equiv (\mathbf{x} = \{x_1, x_2, ..., x_n\}) \tag{3.5}$$

where each x_j represents one feature j of the document, such as the occurrence of a certain word. The derivation of $\mathbf{x}$ from a document is a preprocessing issue that is of the most crucial importance. With respect to this work, it must rely on knowledge gained from what has been concluded from Chapter 2.

According to a taxonomy of classification expressed by Karen Sparck Jones in the 1970s and referred to by van Rijsbergen (1979, Chap. 3), there is a distinction to be made between *monothetic* and *polythetic* classification. Any algorithmic model for document classification requires determined features for similarity measures[3], and if

[3] This is true even for human classification, though much more difficult to pinpoint

the n features of any member of a class are by necessity identical, not only similar, to the features of any other member (or a predefined set of features characterising the class), we have a monothetic classification. Otherwise we have a polythetic classification. Monothetic classification is rare when we are not dealing with "data retrieval", as pointed out by van Rijsbergen (1979), in which exact matches between a user query and the data in the system is the ideal.

The view on documents as being equivalent to sets of continuous-valued features points clearly in the direction of polythetic classification, since it does not seem meaningful to assume classes of documents whose values in $\mathbf{x}$ are all the same. If they are in fact realized as real-valued word frequencies, it is unlikely that the classification would not result in as many classes as there are documents (i.e. $|\mathcal{C}| = |\mathcal{D}|$). So, the roles of these features are to assist in polythetic classification to varying degrees, depending on which class is considered.

Except for the more precise elaboration on how to derive a set of features for a document and what kind of similarity measures (or other bases for classification) is to be applied, classification has hitherto been approximately defined. What is now to consider are the ways in which functional classifiers can be designed and how sets of classes can be defined.

3.2 Modelling the classification task

As has been made clear in the previous section, algorithms for classification always rely on a set of feature-values that have been derived, and which represent the objects to be classified. How to arrive at such representations is highly task-dependent and an issue that will be treated later on, mainly in connection to the experimental setup. In the following section, it is assumed that such features have been derived for the documents to be classified.

This section will consider a few issues that concern general models for classification, thus applicable for most particular classification

in detail.

tasks and not only for the classification of documents along genre dimensions.

As has been pointed out in the introduction to this chapter, classification models are often derived from the area of machine learning, and the topmost subdivision of different algorithms is often based on a distinction between supervised and unsupervised learning. Applied to classification, this distinction becomes one of supervised and unsupervised classification, where the latter is often termed *clustering*. The word "learning" indicates that what is dealt with is in fact a kind of learning process, proceeding from instances of data from which an algorithm induces its classification capacity. To a large extent, classification algorithms lean on empirical evidence.

In some cases, it is worth considering algorithms that implement hand-tailored classification rules based solely on theoretical knowledge. For the kind of tasks that this work is concerned with, such algorithms are less plausible, as thoroughly established knowledge on what characterises documents within different genres is scarce.

3.2.1 Supervised classification models

Every supervised classification model relies on a large set of what is termed "training data" for its implementation. Not only is each document represented as a set $\mathbf{x}$ of feature-values, but actually as a pair of this set and a label c, $\langle \mathbf{x}, c \rangle$, in which c is usually a label assigned by a human classifier, respresenting what has to be considered the one and only correct class.[4] So, classification relies on the assumption that there is a correct classification for each document, even though such an assumption may be theoretically unjustified.[5] Furthermore, there are two other important restrictions in most implementations. These restrictions, however, can be overcome in different ways.

- There is one and only one correct classification. Disjoint classi-

[4] Computationally, $\mathbf{x}$ is rather to be considered an array, which is why the glyph $\mathbf{x}$ is used, and not e. g. X

[5] Section 3.3 will give an account of this inherent epistemological problem of classification applied in libraries. This assumption must also take account of the issues around classification or indexing consistency recapitulated in Section 3.6.1.

fication (Eq. 3.2) is thus mostly assumed.

- In supervised classification it is assumed, in the training (or learning) stage, that c is never absent, because if it were, that instance would be of no use to the algorithm. Exhaustive classification, as defined in Equation 3.3 on page 71, is thus fostered.[6]

If a classifier (of the disjoint type) is defined as any prepared algorithm that takes as input an instance $\mathbf{x}$ and returns a class c, learning a supervised classifier relies on two stages.

1. A set of training data ($\mathcal{X}$), in which each element is a pair $\langle \mathbf{x}, c \rangle$, is given to the algorithm.

2. The algorithm uses the value of c for each instance in conjunction with the feature-values of its $\mathbf{x}$ to arrive at a target classifier.

Algorithms differ in the ways in which these stages are configured. For instance, a Naïve Bayes algorithm calculates probabilities for a document being classified as c, a document with feature-values $\mathbf{x}$ classified as c, and for the feature-values $\mathbf{x}$ in the training data $\mathcal{X}$; a memory-based algorithm stores each pair in $\mathcal{X}$ in a way that makes it efficient to search through the space of instances for the most similar instances; a decision tree learner transforms the data into a sequence of decision nodes, where decisions are based on individual feature-value thresholds induced from the data in $\mathcal{X}$.

A decision tree learner and a Naïve Bayesian learner can be generally described as *abstracting* from the training data. In the case of the decision tree learner, such abstractions may be perfectly viable for a human mind to read as a set of induced rules. A Naïve Bayes learner could be read as a catalog of different probabilities that are fixed and induced from the training data. These abstracting algorithms are also referred to as *eager*, as opposed to *lazy*, algorithms (see below). One advantage with the category of eager algorithms is that they result in compact and descriptive abstractions.

[6] The addition of a pseudo-class labelled ’unclassified’, or something similar, doesn’t effect this restriction.

In such cases it becomes evident that the choice of training data is a very important issue. The training data must be representatively drawn from a sampling frame defined with respect to the kind of collections on which the algorithm is intended to be applied. It also has to be sufficently large to cover possible variations in real world settings.

Of course, what may seem evident here and is worth emphasizing, is that if the classification task is relatively trivial, such as when carefully chosen subject terms are given in highly structured XML documents, rules can be tailored by hand, but such tasks are very time-consuming and probably error-sensitive.

The opposite of abstracting algorithms are those that are said to be built on generalizations, the so-called memory-based or *lazy* learners (Mitchell, 1997, chap. 8). Memory-based learners only store the data in a manner that makes it possible to search through the data in an efficient way.

Another important distinction between different supervised learning algorithms is how they decide on classifications, when given a previously unknown instance.

In the case of decision tree learners, the features of an unknown instance are examined one after another in the order that the tree is configured from the root and upwards. The trees are simply sequences of rules. In the case of Naïve Bayesian classifiers, the algorithm might decide on the class that yields the maximum posterior probability for the particular feature-value set in question: $\mathbf{x}_i \in c_k \leftarrow \mathrm{argmax}_i\, P(c_k|\mathbf{x}_i)$. Regression-based models try to find a path through the space of feature-values that most accurately separates the data according to the predefined classifications, resulting in functions to which an unknown object can be passed on. Memory-based learners search through the data in order to find the most similar instances with respect to the unclassified one: $\mathbf{x}_i \in c_k \leftarrow \mathrm{argmin}_j\, \Delta(\mathbf{x}_i, (\mathbf{x}_j \in c_k))$.

Different memory-based learners also illustrate the difference between what has been termed *local* and *global* classification decisions (Mitchell, 1997, p. 234). In their simplest implementation they are purely local. When deciding on a class assignment it is only the closest examples that are considered. Other implementations, on the other hand, may perform an analysis of features beforehand in order to weight the impact of different features. These are considered to be global methods, because the complete set of training data has an effect on each class decision.

The different models briefly mentioned above are primarily disposed to perform disjoint and exhaustive classification. If overlapping or non-exhaustive classification is sought, they have to be adjusted in different ways. Some models are also originally disposed to treat only binary classification problems, but are commonly combined into a sequence of binary classifiers (cf. Sebastiani, 2005, p. 112f). This is how *Support Vector Machines* have been extended to fit non-binary classification, one of the most popular algorithms today, as they generally turn out to be superior in terms of performance measures for most tasks (see Meyer zu Eissen & Stein, 2004, for a comparison within genre identification).

The size of the training data is important, and as the data needs to be manually annotated this may seem a large problem for real world implementation. However, this problem may be overcome by adding a feedback function to real world implementations, in which a user may be prompted to accept or reject a proposed classification — a function that can be seen in Microsoft's *Hotmail* junk mail filter. Junk mail filtering software in many cases relies on such procedures.

Essentially, one may compare supervised classification, as presented here, to LIS educational settings, where students learn classification schemes by means of class assignment exercises in which the correct solutions are given afterwards by a teacher. A well performing trained human classifier could be seen as the outcome of such training, as will a well performing algorithmic classifier. The larger the amount of training, the better classifiers we get. However, there are certainly exceptions to such a categorical statement, for instance with respect to what is termed *overfitting* or overtraining, when the training data contains noise, i.e. instances that are atypical for the class to which it has been assigned (cf. Sebastiani, 2002, p. 15).

If there is such a thing as supervised learning it is not surprising that there is also a set of models concerned with unsupervised learning. This kind of classification is better termed "clustering" and will be reviewed below. In fact, the term "learning" becomes a bit blurred when applied to clustering, but it is common to refer to clustering as unsupervised learning, which both Manning & Schütze (1999, p. 232) and Alpaydin (2004, p. 10) do.

3.2.2 Unsupervised classification models

The training data we have in unsupervised learning (even though training data is not really an adequate word here) only consist of the features for each instance to be clustered. There is (typically) no knowledge of any predefined classes at all and its purpose in research settings is rather explorative than related to the accomplishment of some task, as the main objective is to find regular patterns in the data. Unsupervised learning (or clustering) is widely used in *data mining* (Mirkin, 2005), and in bibliometrics (e.g. Schneider, 2004, Jarneving, 2005).

Clustering has more qualitative advantages for research and real life applications. The features to be investigated are usually given and the focus is more on the interpretations of the resulting clusters. The feature selection process, so central for research in supervised classification, is more of a matter of preprocessing.

In the view of Mirkin (2005), clustering involves five stages: 1) the choice of data (or documents, if documents are to be clustered), 2) preprocessing of the data, which includes feature derivation and normalization of their values, 3) the actual clustering process, 4) the interpretation of the clusters, and 5) any conclusions that can be drawn from this interpretative activity.

Clustering aims at revealing previously unknown structures in the data, which would, in a sense, be advantageuous and theroretically well motivated for classification along genre dimensions when genres are defined in the sociotechnical way of this work. However, we may run into problems if we consider coarse-grained genres in which artefacts are differently realized. If features are based on e.g. the logical document structure and linguistic expressions related to text type categories, and the document collection to be clustered covers research articles within any kind of subject domain, it is not unlikely that documents on a humanistically oriented topic and documents on a mathematically oriented topic will unwantedly form two different clusters. In any case, clustering typically lets the data, so to say, speak for themselves without the often valuable feature for learning that a manual class label would provide.

Models for clustering are not as many as for supervised classification. They are mainly divided into two groups, hierarchical and non-hierarchical (or flat) models (Manning & Schütze, 1999). One family of hierarchical models tries to break down the collection treewise, beginning with one single cluster that is split into two and so on until all objects are in their own singleton cluster, or until a certain threshold of similarity is reached. Another family of hierarchical models proceeds the other way around beginning from singleton clusters that are merged pair by pair until all objects reside in one cluster, or a threshold is reached. Flat models are all models that are not hierarchical, of which the most well-known are the *K-means* and the *Expectation Minimization* (EM) algorithms.

With an analogy to human classification, unsupervised learning may be compared to an educational setting where students are told to group a set of documents into piles, without any other explicit guidance from the teacher's part. Many introductory textbooks on classification within LIS often begin with sketching up such a scenario for classification, indicating what happens when different humans take different principles for subdivision as a departure for the subdvision of a pile of books. See for instance Broughton (2004) and Hunter (2002). Clustering is close to what Ranganathan (1994, p. 15ff), from a cognitive perspective, refers to as "the first sense" of classification, which in an evolutionary sense is said to precede the individual's capability of classification "with a notation".

3.2.3 Concluding remarks on classification models

There are many classification models to choose from — and many of these are available as ready-made software packages that may stand alone or be integrated into other more encompassing applications. For classification research purposes these packages may easily be used as they are, thereby reducing the efforts to the important tasks, instead of having to write new implementations. The important tasks are 1) feature derivation and selection, 2) choosing or compiling training data, 3) deciding upon a set of target classes, and 4) evaluation. However, being dependent on existing implementations also puts restrictions on what one may do. For instance, most implementations are inclined to target disjoint and exhaustive classification, which means that experiments primarily have to be adjusted to these limitations. This is a drawback to comply with in this work as well.

3.3 Classification in libraries

The account of classification models given in the preceding section risks being considered incomplete, unless we also consider how the process of classification has been treated and studied within LIS. It would certainly be strange to not take account of the more than 100-years-old tradition of classification in libraries, even if it would appear irrelevant with respect to the research questions or turn out not to have much to contribute to the development of algorithms for the classification of documents along genre dimensions. The following section will therefore try to summarize this tradition and thus pinpoint some of the differences between algorithmic classification and human classification.

3.3.1 Contrasting human classification with algorithms

It is not uncommon to find statements in the LIS literature that contrast algorithms for indexing or classification against classification or indexing performed by a human mind.[7] In the "Lifeboat for Knowledge Organisation" (Hjørland, 2006), it is said that if human agents were taught to examine a document in a rigid way according to precise rules, human indexing would essentially be no different from when algorithms are applied for indexing. Lancaster (1998, p. 67) makes a similar observation when reflecting on how human indexers would

[7] As has been declared in the Introduction, the difference between indexing and classification is only superficial and what holds for the task of indexing may also well hold for the task of classification.

behave if they were instructed to assign words and phrases from the document's text only, to come up with a set of "terms" that can be considered descriptive for its contents.

These two observations contrast the human analysis of documents with an algorithm's performance, suggesting that the difference lies in the fact that human analysis is a far more intricate process. Sometimes it is emphasized that human analysis is superior to an algorithm's analysis. In a long footnote Hjørland (1997, p. 51) states: "Real subject-relatedness does not depend on perceptions of similarity but on theoretical analysis! No advanced ... classification can therefore be based on common properties or similarities..." and he proposes the alternative to similarity of "functional equivalence (or isomorphism)". This seems to be a refutal of the otherwise commonly accepted view on classification as building on similarity principles for subdivision and is rather elusive with respect to what is intended with the difference between similarity and equivalence. In one way, he seems to argue for a monothetic classification. However, what Hjørland seems to be making the point of, besides showing a distrust in algorithms, is probably that the analysis of a document on the one hand is not solely context free, and on the other hand that function is probably more important than what a document on the surface seems to be about.

Traditionally, the process of classification in libraries is often said to start with an intellectual effort referred to as *subject analysis*[8] (see for instance Langridge, 1989, Buchanan, 1979, Batley, 2005, Hjørland, 1997, Foskett, 1996). Subject analysis serves the purpose of establishing *subject access points* to a collection, where "subject" must be understood in its broader sense of e.g. topic, discipline, and document or publication type.

Usually subject analysis is said to be followed by a process of translation (Langridge, 1989, Ranganathan, 1960, p. 6, pp. 1-5), where the analysis is mapped to the controlled vocabulary of a classification scheme, thesaurus or subject heading list (see also Hjørland, 1997, p. 44).[9] Such vocabularies employ certain notation schemes and the

[8] There are several other terms used in the same sense as subject analysis (see for instance Hjørland, 1997, p. 39).

[9] It should be mentioned that ISO standard 5963:1985 *Methods for Examining*

process of classification becomes more of a description than a classification. Buckland (2007) and Robert Fairthorne terms this process "marking", and recognize it as an important descriptive issue in which one has to take account for problems related to expectances and conventions among users.

One set of guidelines for subject analysis given by Taylor (2004, p. 347f) to students in library education expresses this translation task. Without considering any restrictive vocabulary, the student is instructed to chose the most fitting concepts (including non-topical ones), combine them to a statement of the contents of the document and then to check this against the vocabulary in order to translate the statement.

In the introduction to the Dewey Decimal Classification scheme (Comaromi et al., 1989, p. xxix), the classifier is directed to particular places in the book for the subject analysis. This, in fact, expresses an embryo for algorithm development. Human classification experience has attained the knowledge that title pages, tables of contents, and similar parts of a book usually convey most of the contents. However, levels of routinization are not particularly illustrative for the essential differences between human classification and algorithmic classification.

In this respect, subject analysis may be looked at as some kind of preprocessing, analogous to the feature derivation process of algorithmic implementations. Unfortunately, these common ways of contrasting human analyses to algorithms, by putting human interpretation next to feature extraction rules, tend to hide several important aspects regarding algorithmic means of providing information access. First of all, what can be mistaken for the whole algorithmic process is only the feature derivation process, or even the mere feature extraction process. As was pointed out in the previous section (Sec. 3.2.1), the feature derivation process is almost always followed by further

Documents, Determining Their Subjects, and Selecting Indexing Terms depicts this as a threefold activity where determining "subject content" is distinguished from the choice of the "most important concepts". This reflects the view of Ingetraut Dahlberg (1978, p. 9), who distinguishes between an ideational, verbal and notational level in classification — distinctions introduced by Ranganathan (1967).

processing, which can be rather advanced and complex.

3.3.2 Descriptive versus classificatory purposes

Probably unwantedly, the observations in the “Lifeboat” and by Lancaster illustrate another difference between the classification in libraries and algorithmic classification. Classification in libraries, and indexing in particular, departs from a descriptive goal, to create a set of terms or symbols that are significant for the contents of the document, not primarily to end up with a suitable subdivision of a particular collection. Indexing in particular, but classification as well, stresses *the assignment of labels to documents rather than the assignment of documents to classes of documents*. This tension between description and subdivision is related to what kind of information access support is highlighted — searching or browsing — which will be further treated in Section 3.5.

It may seem that given a descriptive task or a subdivision task would not matter, the results should be the same. However, this would require that there is one and only one true way of classifying each document in a collection, regardless of the context, situation, and audience. This is probably where the contribution of Hjørland’s research to LIS has been the most influential. His point is that the “aboutness” of a document is not given a priori and that how to characterise the contents of documents has to be domain dependent. This implies that the scope of a collection of documents is influential for every classification decision. There is, for instance, no point in creating a class of which the document members treat economy when the entire collection is compiled for studies in economy, whereas in a general collection of a public library a class for books on economy is indispensable. Thus, a human classifier that is unaware of the scope of a collection would probably end up with a different classification decision than someone who is aware of the complete contents of a collection. Empirical research seem to lack the influence of scope awareness in classification.

When the act of deciding on subdivision principles is detached from the act of performing classification, the human classifier restricts the task to the twofold activity of subject analysis and translation and is not directly concerned with the subdivision of a document collection — subdivision becomes more of an outcome of the local processes of classification decisions made by the human classifiers. Ranganathan's proposal to name the one who designs classification schemes a "classificationist"[10] and the one who applies it a "classifier" (Ranganathan, 1991, p. 11) is illuminating. Subdivision is reduced to a possibility provided by the classification, that is, the subject analysis and the translation. It is the classificationist's decisions when designing schemes that is responsible for the class formations of a document collection.

Indexing and classification can thus additionally be seen as local methods, as opposed to global ones, in the sense discussed on page 78. This is because the subject analysis, and the following translation into terms or symbols from a controlled vocabulary (a thesaurus, subject heading list, or classification scheme), is usually performed without taking account of the entire collection, except when the concept of *literary warrant* (Foskett, 1996, p. 28) is extensively applied, as in the case of the Library of Congress Subject Headings:

> The number and specificity of subject headings included in the Subject Authority File . . . are determined by the nature and scope of the Library of Congress collections. (Chan, 1990, Sec. 3.2)

Algorithms usually take account of an entire collection of documents and must in that sense be regarded as being global. In fact, this is probably a more striking difference between the classification in libraries and algorithmic approaches to classification, than that between human interpretation and strict rules.

This difference can be further illustrated by Figures 3.3 and 3.4. If we think of classes as areas with certain shapes that reflect how

[10] This term, classificationist, thus denotes an individual or a working group of some kind that either elaborates on the principles of classifying library material from a theoretical point of view, or who takes on the role of participant in the design and revision of a classification scheme. A *classifier* is then someone or something that makes use of a scheme in classifying documents.

they embed documents in a two-dimensional space, classification in libraries may be understood as in Figure 3.3, where a document d_1 is drawn into the predefined area of class C_2. The shape of class C_2 is not affected in any way. Contrarily, in Figure 3.4, the shape of class C_2 is being transformed as document d_1 is taken account of, together with all the other documents (this new shape is indicated by a dashed outline). In this case it is assumed that d_3 was the document most similar to d_1, which was embedded in the realms of class C_2. Had it been d_6 that was considered the most similar one, the result would have been a transformation of the shape of class C_1 instead. Figure 3.4 illustrates a completely global classification, independent of any predefined classes. It would thereby also be described as an inductive classification, as apposed to a deductive classification. This, in fact, is similar to what is taking place if clustering is applied.

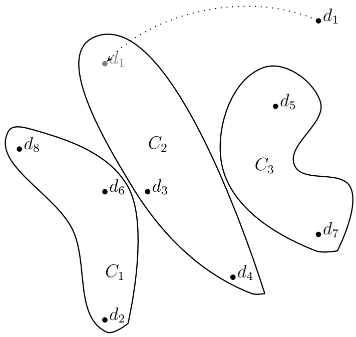

Figure 3.3: Classification in libraries — a local and deductive assignment

However, a global and purely inductive approach would be impossible to realize, because there is at least a need to decide upon some assumptions for determining the resulting number of classes and choosing similarity measures. Real implementations are never completely inductive, even though they may be global. As for human classification, it may be pointed out that the classifier's experience of previously clas-

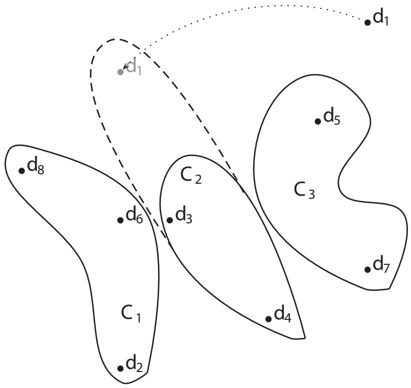

Figure 3.4: A global and inductive classification

sified documents forms a kind of global frame of reference for future classification decisions, so in a sense, a human classifier does not perform a classification completely independently of his or her previous class decisions.

3.3.3 The intendend use of schemes

We have seen that supervised classification generally assumes that classification is performed as a disjoint classification — each document is assigned to one and only one class. In libraries, the issue does not have a straightforward solution. It is here that we may find a motivation for the distinction between the two concepts ’indexing’ and ’classification’ in libraries. Indexing is sometimes a highly overlapping classification task. In many cases, especially in the context of commercial database indexing, an indexer (or classifier) is recommended to be as *exhaustive* as possible when assigning terms to a document. In this case, it must be emphasized, exhaustivity means something different from when exhaustive classification was defined in Equation 3.3. Levels of exhaustivity concern the extent to which the contents of a document is completely described by the set of terms assigned to it. (cf. Taylor, 2004, p. 250f.) This is, again, an emphasis on the descriptive purpose.

However, when library practitioners refer to classification, overlapping classification is termed cross-classification and should be avoided as far as possible. The introduction to DDC is particularly occupied with the renouncement of cross-classification to the point that the motivation for some of its guidelines becomes difficult to make sense of. For instance: "If two subjects receive equal treatment, and are not used to introduce or explain one another, class the number whose number comes first in the DDC Schedules" (Comaromi et al., 1989, p. xxxi). In practice, this results in the somewhat strange fact that the way in which the classes of a classification scheme are enumerated directly reflects classification decisions.

However, even if cross-classification is not recommended for use of the large and general library classification schemes, it has nevertheless become a frequent solution to situations where the classification scheme fail to meet local requirements. (Cf. Svenonius, 2000, p. 112f)

3.3.4 The interdependency between human classification and algorithms

The contrast between human and algorithmic classification can be misleading for another reason. If we exclude clustering, algorithmic classification is never automatic in the sense that it is performed independently of human classification. On the contrary, as was explained in the previous section, algorithms need at least some training data to rely on, and these training data are the outcome of human classification. If the training data contains many questionable or incorrect class assignments, it is highly disadvantageous for the performance of algorithms. Moreover, when developing an algorithm, the algorithm is always implemented by means of constant evaluation against how humans have classified each instance. This evaluation process assumes that humans do it correctly and the accuracy is measured according to this assumption. Last, but not least, algorithms are the outcome of how humans have modelled the task of classification and implemented it as algorithms.

This interdependency is precisely why there is no point in judging human classification as superior or inferior to algorithms, since the latter ones are always imbued with theories of human classification.

3.3.5 Concluding remarks on classification in libraries

Contrasting classification performed by humans with algorithms is possible, and is frequently being done. However, given that algorithms are highly dependent on human classification, such comparisons are not particularly illuminating. What is more important is that irrespectively of the existence of algorithms, is probably more relevant to highlight such aspects of human classification as its local and deductive character.

Every classifier, human or a machine, needs to rely on experience. It takes a lot of practice for a human mind to get to know a classification scheme and to gain efficiency in the subject analysis and translation processes. In this way, the more documents a classifier has seen and been trained on, the better performance he attains. This is also a fact with respect to an algorithmic classifier. The more documents an algorithm has encountered and classified correctly (with respect to a human measure), the better the resulting classifier usually performs, if this learning phase is appropriately configured — "better" from a human expert's point of view, that is to say.

It is not easy to say what the task of classification in libraries can contribute to the development of algorithms or, for that matter, what it actually have contributed to the existing models that are mostly developed with other tasks in mind than the classification of documents. However, even though the knowledge of classification in libraries to a large extent eludes description, the outcome of human classification is mostly regarded as a "gold standard" for the development of algorithms, and can therefore not be ignored. Classification in libraries probably deserves the most attention for its emphasis on the design of classification schemes, which will be the focus in the following section.

3.4 Defining a set of classes

Most experimental work on algorithmic classification proceeds from a given set of classes to which objects are to be mapped. The problem in itself, to which classification is a methodological response, often implies an assumption of a finite set of predefined target classes, which is more or less well defined. For instance, information retrieval research considers two classes — relevant and irrelevant documents; junk mail filtering is supposed to weed out junk mail; parts-of-speech tagging (a linguistic problem) maps word occcurences in a closed text to a finite number of word classes. Even though the precise definition of the classes in these classification problems presents difficulties, the classes are at least loosely given a priori in terms of their labels — {junk, non-junk}, {relevant,irrelevant}, {noun, adjective, ..., pronoun}. Algorithmic classification assumes this "given-ness" since what is in focus is the development of algorithms or feature sets that perform well, not the exploration of possible target spaces of genres.[11] If the goal is to increase performance, experimental work needs to rely on available preclassified data, to measure any improvement of the algorithms. Hence, the relatively few classes are given a priori and the objectives are seldom to refine the set of classes in existing corpora. The task of defining sets of classes is thus relatively often ignored in the context of algorithm development.

It is only recently that the more elaborated development of classification schemes for genres has been given special attention in the context of experimental research on classification along genre dimensions, especially within the WebGenreWiki-community (see p. 64) and at the Colloquium *Towards a Reference Corpus of Web Genres* held in Birmingham, UK, in 2007. Two issues are shown to be of special importance here, except the obvious need for preannotated document collections (i.e. corpora). First, the set of target genre classes

[11] It is worth emphasizing that clustering is often mentioned as a process which may appropriately prepare for classification, where the patterns of regularities in the data that are revealed may also enhance classification performance (Alpaydin, 2004, p. 145). See also how Santini (2005) used clustering for explorative purposes in genre identification.

needs to be warranted, e.g. with respect to user expectancy. Second, any potential relationships between different genres may be theoretically and empirically investigated.

However, in addition to defining classes, it is also possible to look for existing classification schemes that may in whole or partly incorporate classes related to genres. This is the case where the long tradition of classification in libraries comes in handy, as this tradition really is mostly occupied with the design of controlled vocabularies (or more specifically, classification schemes).

3.4.1 Classification schemes

In Section 3.1, the space of target classes of any classification task was referred to by the symbol C, and this space consists of classes referred to as a set $\{c_1, c_2, ..., c_n\}$. When the task of defining C is considered in library contexts, the document collection D intended to be subdivided is usually not considered, at least not in depth. In addition, according to Miksa (1992), the main purpose of classification in libraries before the First World War was also educational, in which the aim was to disseminate a depicted structure of human knowledge by means of a thoroughly devised classification scheme. On the other hand, Melvil Dewey, who initiated the development of the Dewey Decimal Classification system (DDC), illustrates an exception to these somewhat overambitious claims. Frohmann (1994, p. 112) points out that Dewey strongly rejected the idea that his system maps any "structure beyond class symbols". The history of classification schemes within library practices is long and mostly detached from the classification of particular document collections, with one exception. A strong dependency on a collection (expressing *literary warrant*) is, according to Foskett (1996, Chapter 22), demonstrated in how the Library of Congress Classification (LCC) scheme came into being in the early 20th century. The LCC was explicitly developed in order to reflect an appropriate partitioning of the, at that time, current collection of the Library of Congress.

The classes of classification schemes are mostly seen as representations of topics, concepts, phenomena, etc (cf. page 16). In the simplest case, however, C can be defined by one or more classificationists by simply enumerating and describing the characteristics of each $c \in C$, one after another — a simple list of "subject headings". Such an enumeration can be accompanied by guidelines for the application of each class, that is, more or less strict rules for the inclusion of documents into classes when applied to documents of any kind.

However, choosing a topical example for now, when C contains two classes, where one is for books on dogs and the other for books on wolves, the two concepts represented by the classes are related to each other in a generic fashion. In scientific classification, `canis lupus familiaris` is a subspecies of `canis lupus`, which in its turn is a subspecies of the `canis` genus. Therefore dogs are generically subordinated wolves. Subordination and superordination can be, and are extensively, introduced in library classification schemes and thereby also impose a hierarchical, tree-like (or "genealogical") order upon the target classes.

Library classification schemes are not scientific classification schemes, but they mimic similar structures. So, C is usually not defined as a flat set of classes, but as a class of subclasses of subsubclasses, etc. If we denote the level of each class with a superscripted index, as in c^0 for the root level and c^1 for its offspring level (subordinated, that is), we may have a structure of classes where

$C = \{c^0_1, c^0_2, ..., c^0_n\}$ `and, e.g.,`
$c^0_1 = \{c^1_1, c^1_2, ..., c^1_n\}$ `...`

So in the Dewey Decimal Classification (DDC) we have, for instance, as c^0_2, `Philosophy`, subdivided into `Metaphysics`, c^1_1, and `Epistemology`, c^1_2, etc. All the while it should be noted that all classes on level 0 may not only have other classes as members, but are also intended to be used as containers of documents themselves, from a set theoretic point of view. Thus, from the point of view of the classification scheme, documents assigned to the class `Epistemology` can be seen as contained in the class `Philosophy` as well, and the result is then not really a disjoint classification at all.

The scientific claims remain though, imposing the semantic web of lexical units upon document collections. In an introduction to the DDC scheme, there is an explicit reference to the ten main (level 0) classes of the DDC scheme as something which "together cover the entire world of knowledge" (OCLC, 2003, p. 2). Sue Batley (2005, p. 3) states that besides organizing books, library classification schemes "list the main and subsidiary branches of knowledge, so they provide a taxonomy of knowledge...". These are large claims and draw attention to the (sociological and epistemological) criticism of classification schemes, which has become more and more frequent in recent years (Hansson, 1999, Samuelsson, 2008, e.g.).

Hansson (1999) has, by reading the Swedish SAB classification scheme as a text in its own right, made a thorough critical investigation of how the design of the classification scheme reflects a certain society's view on the world. For instance, he makes the observation that the top class 'religion' (denoted by `C`) is divided into a set of 13 subclasses of which only one is attributed to non-Christian religions. He also claims that, in many cases, the design of classification schemes is approached as a kind of knowledge representation that reflects certain epistemological views (Hansson, 1999, e.g. pp. 31-32). Birger Hjørland (1997) makes an even stronger point on the impact of differing epistemological foundations on LIS research in general.

Observations similar to that of Hansson's have inspired the domain analytical approach of Birger Hjørland, in which the idea of the so-called universal classification schemes is rejected.[12] Classification schemes, according to Hjørland, have to be designed with a fairly narrow domain in mind, representing the "knowledge structures" within that domain (Hjørland, 2002, cf. e.g.). Ørom (2003) gives a fairly recent account of how the domain of Art Studies requires specialized structures of knowledge representation.

Usually, the hierarchical nature of classification schemes does not necessarily have any effect on the classification task itself, especially not when algorithms are concerned. Ehen mapping documents to classes, it does not matter for the classifier if the class of dogs is related to the class of wolves. This is, however, highly relevant for

[12] I.e. schemes that encompass any conceivable topic or phenomenon in any context

the human classifier when “translation” is to take place, and the investigation of the classification scheme for finding the correct class is undertaken. Thus, the hierarchical nature becomes a valuable structure as the scheme is presented to the classifier — a matter of display and visualization. Jacob (2004), e.g., dismisses non-hierarchical schemes (i.e. flat enumerations of classes — subject heading lists) as mere “categorization” tools. Most of the theory of library classification is concerned with the definition of such structures — and how to construct labels for the classes in order to show their position in the hierarchy.

There is one more concern of classificationists that is important to highlight, regarding these labels. In the description of hierarchies above, a certain set theoretical notation was temporarily used, which is not adequate for human classifiers. Instead, many classification schemes employ a certain expressive notation that indicates at which level of the hierarchy a certain class resides. For instance, simplifying a fairly complex and intricate matter quite a bit, in the DDC `Philosophy` is denoted by the symbol `100`, `Metaphysics` by the symbol `110`; and in the Swedish SAB system `Philosophy` is denoted by the symbol `D` and `Metaphysics` by the symbol `Dj`. In the DDC, the number in the second position refers to a subclass, when it is not zero; and in the SAB scheme, the addition of a letter after `D` signifies that it refers to a subclass of `D`.

These examples illustrate the fact that the two schemes have different *bases of notation*, that is, the range of symbols applied for one level in the hierarchy. The DDC uses the symbols $\{0, ..., 9\}$ and the SAB system letters from the (Swedish) alphabet (even though a few of them are omitted). How to combine these symbols in order to reflect which class is referred to is a complex issue which falls completely out of the scope of this work, as it deals with the display of classification schemes and not the task of classification. What is important to note is however that the choice of notation base generally restricts the number of classes on a certain hierarchical level, in a fairly haphazard way. There is no reason outside the pragmatic context of library classification that motivates a division of topics, disciplines or anything into a number of main classes restricted to the cardinality of the set of symbols used.[13]

The long tradition of library classification conveys another important matter that is worth considering. This matter concerns the elements of a classification scheme and their nature — or rather the nature of the phenomena and concepts they refer to. In the examples given thus far in this section, classes refer to academic disciplines or domains of studies, but classes may refer to complex and compound topics as well, which are more difficult to position within a strict hierarchy.

The *Library of Congress Classification scheme* (LCC), originally designed towards the end of the 19th century, and still in use – although much revised, was defined as a closed finite space of classes, and has therefore been characterised as an *enumerative* scheme. Ranganathan's *Colon Classification* (CC), with its first edition published in 1933, was more flexible and did not enumerate classes, but rather it provided principles for the construction of class labels based on numbers and interpunctuation. This classification scheme is an eminent example of how the subjects of the class numbers can be broken down into analytic constituents. Ranganathan himself gives an example of how the class `Treatment of tuberculosis of the lungs by X-ray` is formed by facets within the "group" Medicine (Ranganathan, 1991, p. 57).

```
Medicine [Lungs]:[Tuberculosis]:[X-ray treatment]
```

Here the names within square brackets are instances of the facets `Organ`, `Problem`, and `Handling`, respectively.

In a facetted scheme like this, a set of predefined facets are pulled together in order to form a "class", which is essentially described by the facets themselves. This concatenation is made by the classifier: "by combining the numbers in the different unit-schedules in assigned permutations and combinations, the Class Numbers for all subjects can be constructed" (Ranganathan, 1960, p. 12f). Most universal clas-

[13] Of course, there are several ways to overcome such a restriction. Ranganathan introduced the concept of an "emptying digit" and there are attempts to use "octadecimal" numbers (where the symbols $\{0, ..., 9, A, ..., H\}$ are used).

sification schemes today have also integrated this feature.

Expressed somewhat more freely, classes are defined by an analysis in different dimensions, so that the classifier is actually partly responsible for the formation of a classification scheme, even though the dimensions are given beforehand. From the algorithmic point of view, such tasks may be approached by means of parallel classifiers, each trained on and using different sets of features — at least in theory.

For Ranganathan, and many others within the tradition of knowledge organisation, the notational conventions of elements in a scheme are central phenomena, as are the relationships between elements (cf. Dahlberg, 1978). This tradition forefronts classification schemes as knowledge representation tools.

The notation of a classification scheme has been considered something that falls out of the scope for this work. This marginalization does not necessary include the notion of the naming of classes. The naming or labelling problem is given a recent and short overview by Buckland (2007), where he emphasizes this as an issue that requires both backward-looking and forward-looking. The name must "be derived from the discourse from which the name originates" as well as respect the fact that it is for future use and "ways of thinking". With respect to genres, one may consider how the term "weblog" has evolved into "blog". Few people searching for blogs would think of searching for weblogs as well. The naming of genres is a complex and crucial issue, intimately related to certain types of information seeking tasks. However, this issue can be separated from the classification task under certain circumstances, which will be clarified in Section 3.5, and is largely ignored in this work.

The idea of introducing hierarchical classification schemes for genres (or classes of documents related to genre) is not new, but still fairly unexplored. Just recently, Stubbe et al. (2007) proposed a scheme of 7 top classes (Journalism – Literature – Information – Documentation – Directories – Communication – Nothing) under which they gather several rather diverse "genres" as subordinated. For instance, below `Documentation` is found, only, `Law`, `Official Report` and `Protocol`. Below `Information` we find, e.g., `Bilingual Dictionary` and `Science Report`. Intuitively, it is hard to understand the rationale for these groupings as long as there are no specific definitions of their applications. The hierarchy is not used for anything other than the arrangement for display of 32 "genres".

To summarize this section on classification schemes, a few words on the classification of classification schemes are worth mentioning.

Hjørland (1997, pp. 46-47), among others, distinguishes between three different approaches to classification that represents different "levels of ambition": 1) ad-hoc classification, 2) pragmatic classification, and 3) scientific classification, where the first is influenced by "private taste" (cf. Taylor, 1999, p 176), and the last one is illustrated by botanic taxonomies. The classification schemes employed consequently reflect these ambitions. For instance, information retrieval relies on ad-hoc classification, beacuse it divides a collection into relevant and irrelevant documents according to user preferences expressed by means of a query. Depending on the scope of the collection or the intended scope of use, schemes can also be characterised on an axis ranging from general to specific. The most general schemes are aimed at organising documents on any conceivable topic and kind, whereas the most specific ones are developed for collections with a restricted space of topics or comprised of particular kinds of documents, such as audio recordings. The scope of use for a particular scheme may also be more or less restricted with respect to nation, language or type of corporate body. Koch et al. (1997) add to this typology the "home grown" schemes (similar to Hjørland's ad-hoc classification), exemplified by the Yahoo categories, but this is a category that seems to depend on the degree of standardization or theoretical foundation, and not primarily its scope of use.

3.4.2 Concluding remarks on defining classes

The development of classification schemes is a fairly unrecognized problem within the tradition of experimental research on algorithmic classification, especially with respect to genres. The long history of classification in libraries indicates that there are several crucial issues for the task of defining classification schemes: 1) the design of a hierarchy, 2) the choice of principles for expressing this hierarchy in a notation, and the naming problem itself, and 3) determining the scope for its use and the level of ambition.

To summarize and conclude, it can be stated that classification schemes that are designed to function in library practices are often fixed constructs that over time become outdated as knowledge representation tools. There are different ways to come to grip with this problem and, as Beghtol (1998) seems to suggest, giving as many views as possible on a document collection may come to grip with a variance in the uers' knowledge structures. .

To consider using existing schemes for algorithmic application is far from realistic. The number of classes is much too large and thus presents a much too high task complexity for algorithm development. Besides, as has been pointed out in Chapter 2, library classification schemes do not only enumerate classes of the same kind, but a mixture of, e.g., topics, academic disciplines and genres, which would require different kinds of feature sets. This would probably make the feature selection process in algorithm implementations for library classification a very difficult, if not impossible, issue.

For topical classification, using the structures defined in classification schemes to support an ordinary algorithm is higly plausible and has been tried out by e.g. Chiang et al. (2007) in a recent project. For the classification of documents along genre dimensions, the use of hyponymic or meronymic relationships between different genres is unexplored, and will probably remain so. Even though such relationships are not clearly spelled out in library classification schemes, where such relationships may be discerned they could be considered, not the least because they can be considered to be fairly established.

In this work, however, such directions for investigation are not pursued because whereas classes related to genres can be clearly discerned in existing classification schemes, they are generally out of the scope for this work.

One has to admit that though classification in libraries puts a lot of concern into the task of designing classification schemes, this is mostly treated as a collection-independent task with a stress on semantic relationships between topics, phenomena and scientific disciplines. Even though attempts to pinpoint relationships between genres have been made, it is hard to see what advantages such attempts may serve with respect to genre classification in the sense of subdivision. However, given a collection classified along genre dimensions, there always remains the question of how to present such a collection in order to provide an access that is convenient for the user. This question is out of the scope for this work, but deserves some recognition because it relates to the question of why classification is performed — the objectives of classification.

3.5 The objectives of classification

It has been somewhat taken for granted that classification is performed in order to support information access. It is hard to question the fact that an organized collection of documents provides better access to its contents than a collection randomly put on shelves or whose document titles are displayed on a computer in no predictable order at all. However, exactly in which ways classification is considered to support information access has not been spelled out.

First of all, classification has been a prominent activity within library practices for centuries and one may start by asking how it is, and has been, motivated within such practices.

Originally, classification was adopted in libraries for shelving books and other material, but also for a similar organization of contents in catalogs, which functioned as indices to the holdings of a library or to all of the published works within a restricted time span. Oddly enough, these catalogs "were thought of as works to read" themselves, in the late 19th century, if we are to understand Miksa (1992, p. 107) correctly. The main principles (objectives, one may say) for the catalog were described at that time by Charles Ammi Cutter, who later came to influence the design of the Library of Congress Classification scheme (Taylor, 2004, p. 60).

1. To enable a person to find a book of which either
   (a) the author
   (b) the title
   (c) the subject is known
2. To show what the library has
   (a) by a given author
   (b) on a given subject
   (c) in a given kind of literature
3. To assist in the choice of a book
   (a) as to its edition (bibliographically)
   (b) as to its character (literary or topical)

(Cutter, 1904)

These principles are often considered largely valid for the library practices today, even though Cutter's phrasings remind us of the more than 100 years that have passed since he wrote the "rules". In practice, classification is thought of as obviously supportive of objective 1 a), 2 b), and probably also 2 c) and 3 b), because as was mentioned in Section 2.1, classes are not restricted to subject based criteria.

Library practices have since the time of Cutter resulted in more elaborated and precise rules for how to proceed with bibliographic descriptions of different kinds of documents, such as the Anglo-American Cataloguing Rules (AACR). The AACR and its intricate rules, together with its standardized forms of representation, MARC, are nowadays given special attention in the education of librarians. Classification is an integral part of this more encompassing *descriptive* activity. Thus, in addition to supporting shelving tasks, classification also supports a descriptive objective.

With Cutter's rules in mind, we may say that the bibliographic record fulfils certain functions for the user of the catalog who has the intention of finding a certain document (rule set 1), getting an overview of the collection (rule set 2), or choosing between candidate documents (rule set 3). The functions of bibliographic records have been (re)articulated recently in an influential work carried out by a working group of the IFLA (International Federation of Library Associations and Institutions), often referred to as the *FRBR* — the acronym for *Functional Requirements for Bibliographic Records*. The explicit purpose of the working group was to "delineate in clearly defined terms the functions performed by the bibliographic record with respect to various media, various applications, and various user needs" (Study Group on the Functional Requirements for Bibliographic Records, 1998, p. 2). The functions defined can be seen as a refinement of Cutter's general principles and the FRBR made explicit the bibliographic record's support in a) finding, b) identifying, c) selecting, and d) obtaining "entities".

In the FRBR, Cutter's 'books' were substituted by 'entities', his rule set 1 and 3 corresponding roughly to functions a) and c) respectively. Functions b) and d) are, compared to Cutter's rules, an addition. This is to a large extent due to the fact that the perspectives of Cutter and FRBR are different. 'Entities', for instance, are not restricted to documents. The FRBR are not only occupied with access to material items (books) but also with access to abstract "works", "expressions", and "manifestations".

However, the FRBR's focus on records instead of collections indicates that the FRBR not seem to be concerned with the important function of providing an overview of the collection, of *visualizing* the collection (of collocation, Cutter's second rule set). Svenonius (2000, p. 17ff) has suggested a set of amendments to the FRBR functions, where she takes account for collocation and navigation support.

It has been pointed out by Svenonius (2000, p. 17ff) that "the collocating objective" and "the navigation objective" of bibliographic description to some extent have become marginalized objectives (cf. p. 102). This is evident both within the context of library focused efforts such as the FRBR and in the context of information retrieval research, where algorithms that match queries with document representations (bibliographic references) and models for automated extraction of index terms have been the predominant issues. It seems that it is often assumed that users tend to favour *querying* before *browsing*, even though there are no clear indications of such an assumption being generally true. Koch et al. (2004) have examined the proportional use of browsing in the Renardus system, that uses the DDC scheme as a browsing aid, and found that browsing was the dominant method of access. Nicholas et al. (2006), on the other hand, have found the opposite tendency in the usage of a library's collection of electronic journals. Both investigations were made by analyzing usage log data.

The notions of querying and browsing reflects two ways of gaining information access that characterizes two sometimes competing objectives for classification, where the former stresses the importance of descriptive power and the second the importance of visualization and navigation. The differences between these two objectives seems not to have gained much attention in LIS classification theory.

Querying should be understood as directly observable interactions between a user and a collection of documents (or document representations) carried out in an interface that algorithmically matches a verbalized expression to the contents of the collection. The matching documents or document representations that are, so to speak, filtered out, are supposed to be relevant with respect to the verbalized expression. They are ordered according to a specific principle (e.g. alphabetic, chronological or according to expected relevance), but are not in general subdivided. Nevertheless, the result of querying could be described as a subdivision of the document collection into two classes, relevant and irrelevant, where the latter class is hidden and unbrowsable, since it is actually discarded.

The collocating and the navigation objective would state that sets of documents should be arranged in a surveyable and conceivable sequence in order to be traversable. Walking through the shelves of a library, where books are visibly ordered according to some principle, is somewhat prototypical of what has been termed browsing.

Most use of collections is a mixture of querying and browsing, and the distinctions above are components of analytical frameworks that tend to be more or less influenced by cognitive perspectives.

In the tradition of information seeking behaviour studies, querying and browsing signify distinctions between two different types of use of an organized collection of documents, which Ellis (1996, p. 144) refers to as different "search strategies". Marchionini (1995) makes a distinction between "analytical search strategies" (covering "querying") and "browsing", where the former term refers to interactions in which the user has more well-defined tasks and understandings of his or her needs and/or the capacities of the knowledge organizing system. Analytical search strategies that involve queries are particularly applied by professionals — librarians or information specialists using commercial databases. The development of such strategies is considered one of the main issues for LIS education, perhaps at the cost of browsing competence. Browsing refers to interactions where the goal is more "informal or general" and puts less cognitive load on the user (Marchionini, 1995, pp. 102-103). It usually relies more on the structure formed by the relationships between elements of, for instance, a classification scheme, which then provides for the traversal of the space of classes within the scheme. Hypertext is sometimes referred to as the prototypical application of systems for browsing.

Search strategies are often analyzed and further divided into subtypes, such as briefsearch, pearl growing, scanning, and exploring strategies (Harter, 1986, Ellis, 1996). Such fine-grained typifications of user interaction applied in the research of information seeking behaviour pay special attention to the user's state of knowledge and are considered important for the design of interfaces. In essence, they are in some sense mentalist, drawing on theories from cognitive psychology. There are less obvious cognitive approaches, as Bodoff (2006, p. 69f) points out, that relate browsing to the observable behaviour of users, independent of any specific task or any information need. However, such approaches have been less successful within LIS. It is tempting to speculate with Bates (2002), that when the collection of documents is topically more constrained, the tendency to browse is increased, whereas querying is more apt for collections with a highly skewed distribution of potentially relevant documents.

Hjørland (1998, p. 20), one of the critics of the *cognitive viewpoint* of LIS, refers to these approaches as subjectivist and considers them inadequate for LIS and for the development of knowledge organising systems in particular. However, a distinction between query support and browsing support is tied to the systems used, and not to how they are used or how the human mind approaches them. Classification schemes can be used to support querying as well as browsing, but when it comes to query support, labels (the notation) are in focus, and when it comes to browsing support the subdivision principles are in focus. The question at hand for this work is the subdivision of document collections, where browsing support is a particularly interesting outcome, and the labels or names of classes are of less interest. The task of providing query support requires careful attention to the labelling, whereas providing browsing support is a kind of visualization that does not necessarily require labels at all. In fact, if the type of classes intended are only vaguely known and intuitively recognized by users, labels may be confusing.

Even though LIS and affiliated areas of study debate the usefulness of different designs of interfaces to document collections and in general tend to favor issues regarding query interfaces, it is evident that both query and browsing support is of high value. However, if we deal with aspects on documents where we cannot rely on the assumption that users know the names of classes, as is the case for some genres, developing support for browsing becomes a crucial necessity.

The stance taken in this work is that as long as genre is a vague concept (possibly, as Rosso (2005, p. 17) – drawing upon Biber – thinks, a "folk typology"), classification along genre dimensions in support of querying is an extremely difficult task, when names of classes are forefronted as access points. Browsing support is much more plausible, since the ways in which the representation of classes will be accessed is not predetermined and need not rely on established terminology.

3.6 Evaluating classification

First of all, classification accuracy in libraries is often judged in terms of the accuracy of the subject analysis and the following translation into a controlled vocabulary. In other words, again, it is not the actual subdivision of a set of documents that is important, but the subject analysis. Accuracy has been interpreted in many ways. For instance, accuracy may be seen as attained if the documents of a particular class are similar (in terms of subject matter) to some (abstract) prototypical exemplar of the class, which is how Buchanan (1979, p. 9ff) seems to regard it. For Langridge (1989, p. 1), on the other hand, accuracy is attained if the nature of documents "is correctly perceived", which implies a very idealistic point of view. Hjørland is generally pragmatic, with his "domain analytic" stance, where descriptions should be made according to the terminology of the (scientific) domain at hand, for which classification is performed.

Any evaluation of classification by means of algorithms should answer questions related to how well a classification task is accomplished. It has been pointed out above that classification accuracy in libraries is sometimes approached in a rather idealistic way (platonic, that is), insofar as there is one single correct way to perform a subject analysis and translate it to a controlled vocabulary. This is also a necessary assumption underlying supervised learning.

An opposing view on the evaluation task is to assume that only the user can judge the accuracy of a classification with respect to his or her information seeking task. The consequence of this view is that there is no ideal way of dividing a set of documents into classes.

The first view runs into problems when we consider the fact that there are lots of empirical evidence for the fact that even experienced librarians do not always agree with every classification decision and that the same librarian may arrive at different content descriptions on different occasions (Lancaster, 1998, p. 62). This issue will be discussed in the next section. The second view runs into problems for the same reasons, people are different and expect different ways of organising documents. This means that the evaluation becomes dependent on who judges the accuracy. In addition, a user centered evaluation makes it difficult to isolate the classification from issues regarding e.g. the interface.

In evaluating the classification performance of an algorithm, one has to take a pragmatic stance regarding these two problems. By far, the most common way to arrive at quantifiable measures of a classification task is to proceed from the assumption that humans who classify a document do this in a correct way. This is the approach chosen for applied research on algorithm development. However, within

LIS some empirical studies on interhuman agreement on indexing and classification have been made, which deserves attention. This is the focus for the next section.

3.6.1 Consistency in human classification

Studies within LIS on the quality of classification are usually referred to as studies on indexing consistency. These studies are usually concerned with tasks where the interest is placed on the choice of terms assigned as descriptors, not on the division of document collections or the assignment of one single descriptor to each document. Additionally, they also concern topical designators. One exception is Subrahmanyam (2006), who studied how LC class numbers were assigned by 52 different libraries. A prominent drawback with this study is that these 52 libraries were all part of a bibliographic network (a bibliographic exchange program) that collaborated on cataloging in such a way that each participating library can hardly be said to have classified each document independently. (See Taylor, 2004, for an account of this framework for *copy cataloging*.)

When indexing consistency is studied, two or more professional indexers are usually told to assign terms to each document (freely, derived from the text, or chosen from a controlled vocabulary), in a way that exhaustively describes the contents of each document. In some cases, the terms are ranked in order of importance. Lancaster (1998, Chapter 5) gives an extensive account of previous studies on indexing consistency, and mentions that in 1965, figures were reported that showed a 24% and 80% agreement, even though it should be remembered that there are several ways to define percentages of agreement in interindexing consistency studies. One more study was made by Leininger (2000). He studied interindexer consistency in the bibliographical database PsycINFO. Even though one has to bear in mind that there are several ways of calculating consistency scores and there may be other factors influencing the results, his figures are discouraging for anyone who trusts human classifiers in evaluating algorithms. The PsycINFO database has a field for the classification codes to which only one (with some exceptions) code should be assigned.

The consistency was estimated as around 45%. However, given that there were 157 different classes to choose from, the task complexity was far larger than in most research on classification algorithms.

When we consider genres, an interesting, but loosely defined, experiment is performed and presented on the WebGenreWiki (2008), where 8 genre researchers are asked to assign genre names to a set of 50 documents. Only with respect to 6 of these documents (12%), do 5 or more researchers agree on the same label assigned to the documents. However, one has to take account of the fact that the experiment does not seem to have been performed with a controlled vocabulary of genre names to rely on, and thus we cannot draw any general conclusions on the task complexity and consistency tendencies for human genre classification.

Another witness on the consistency of classification along genre dimensions is to be found in the comprehensive user study made by Rosso (2005), but in this case the main issue was to establish a space of genres in human cooperation. The agreement measures were a means by which progress could be shown. Santini (2007) also performed a small user study of agreement in disjoint classification of 25 web pages. Only for 5 of these pages an agreement above 80% was attained. In these studies there are no conclusive indications on that genre classification would be more trivial than topical classification.

Anyhow, as has already been pointed out, the low consistency often reported needs to be considered below, as algorithm performance is usually measured with respect to human classification.

3.6.2 Performance measures

Measuring the performance of an algorithm is a question of estimating how good a certain algorithm is for a classification task, or how appropriate a certain set of features is in representing the objects to be classified. Supervised classification assumes a so called “gold standard” for this purpose, which implies not only that a classifier learns from how humans would classify objects, but also that the performance of a classification algorithm is best measured with respect to such a gold standard. For several classification tasks within natural language processing, supervised learning has been used extensively and evalutated against such gold standards (see Jurafsky & Martin, 2000, p. 308, for an explanation of gold standards in natural language processing tasks). In this context, where e.g. the algorithm's task is to determine the correct parts-of-speech in natural language texts, supervised learning was shown to be rather successful when evaluated against carefully compiled and annotated corpora (cf. Megyesi, 2002). By far, human disagreement on to which word class a certain word should be assigned probably tends to be lower than for topical or genre-based document classification, as reported in the previous subsection. In order to properly prepare a valid evaluation of algorithms for classification according to genre adherence, it would be appropriate to carefully compile a set of data tested for human agreement.

Performance measures may play two different roles in research settings. On one hand, an algorithm may be adjusted to perform better based on evaluation measures, and on the other hand, an algorithm (and the feature set) may be judged as more or less appropriate for the classification task at hand (Sebastiani, 2002, cf. p. 12). It is common to refer to the former objective as a process of *tuning* and the latter as a process of *performance estimation* (cf. especially Lavesson, 2006, Chap. 2).

Tuning builds on the premise that any classifier that relies on parameter settings, such as feature weighting, needs to be tuned in order to obtain its optimal settings with respect to the task. Then, given a data set (representing a gold standard) for experiments, these may in fact be divided into three parts. One for training (or learning), a second one for tuning of the classifier, and the last one for estimating the resulting performance. This is a common way to proceed in text categorization research, and is aimed at assuring validity (Sebastiani, 2005, Megyesi, 2002, Manning & Schütze, 1999, p. 114, p. 30, p. 584, respectively). In order to minimize bias it is also important that these three sets are disjoint.

The two processes of tuning and performance estimation require that some subsets of $\mathcal{X}$ (the complete data set) are held out during training. Tuning can be seen as a way of simply evaluating the classification task with respect to the algorithm, as the results measured are not measured independently of the algorithm. They are not algorithm independent, since the variables are set according to the inductive bias of the held out validation set together with the training set. Performance measures that are to be independent of the algorithm, and thus better estimates of the true error that a certain algorithm would result in for unknown objects, have to be derived from a subset of $\mathcal{X}$ that is used in neither training nor tuning.

The measures used for both tuning and performance can be given in terms of *accuracy*, i.e. the correctness of all class assignments, and in terms of *recall*, *precision*, and *F-scores* for each class.

The *accuracy* of the classification is the portion of instances correctly classified. This measure is given in Equation 3.6, where $|X_{correct}|$ is the number of correctly classified instances and $|X|$ is the total number of instances in the test data.

$$Accuracy = \frac{|X_{correct}|}{|X|} \tag{3.6}$$

The result may also be represented as a confusion matrix, as in Figure 3.5. Let us say that we have three classes , $\mathcal{C} = \{c_1, c_2, c_3\}$, and 25 instances to test our algorithm on, $|X| = 25$, and the matrix in Figure 3.5 is given. The accuracy may be determined as $\frac{20}{25} = 0.80$ by summing up the figures on the diagonal from top left to bottom right and divide by the total number of instances. Here, for two instances of class c_3, the classifier has incorrectly assigned the instances to class c_1. This means that the *recall* ϱ for class c_3 is also 0.80, $\frac{8}{10}$, as recall is a measure of the number of identified instances of one class (Eq. 3.7). The measure of *precision* π, on the other hand, is the ratio of correctly classified instances within the predicted class and the total number of estimated assignments for that class. Following Equation 3.8, the precision for class c_3 is $\frac{8}{11} = 0.73$, while both recall and precision are perfect for class c_2 (1.0). $X(c_i)$ in Equation 3.7 and 3.8 refers to the subset of X preclassified as adhering to class c_i.

$$\varrho_{c_i} = \frac{|X(c_i)_{correct}|}{|X(c_i)|} \tag{3.7}$$

$$\pi_{c_i} = \frac{|X(c_i)_{correct}|}{|X(c_i)_{assigned}|} \tag{3.8}$$

```
                c_1      c_2      c_3
           -----------------------
    c_1   |      7        0        3
    c_2   |      0        5        0
    c_3   |      2        0        8
```

Figure 3.5: Confusion matrix

The confusion matrix is a helpful visualization of misclassifications. Recall and precision are valuable measures, as they point to classes that are fairly easy to predict and classes that pose problems.

However, recall and precision have often been shown to depend on each other in a way that an increase in recall results in decreasing precision. It is easy to see that a perfect recall can be attained for a class if all instances are assigned to that class. A perfect precision may be attained if a set of features that are unique for one class within a certain collection are chosen. The first strategy is awkward as it leads to zero recall for any other class, and the second strategy may lead to an overfitting problem, where unknown instances do not fit the algorithm at all — a maximized inductive bias. Because of this it is plausible to use a harmonizing measure called *F-score* to measure the algorithm's performance for each class. If recall and precision are considered to have equal importance, Equation 3.9 is applied, otherwise it can be adjusted in the ways described by Sebastiani (2005), Manning & Schütze (1999, p. 269), originally proposed by van Rijsbergen (1979).

$$F_{c_i} = \frac{2(\pi_{c_i} \times \varrho_{c_i})}{\pi_{c_i} + \varrho_{c_i}} \tag{3.9}$$

However, there are two problems that remain. First, by tuning an algorithm based on a sample we are at risk of overfitting the algorithm to the sample and it may perform badly when confronted with a completely different sample. Second, we need some knowledge of the task complexity in order to determine whether the algorithm performs well or not with respect to the task.

The first problem is inherent to the task of this work. The large corpus of documents for which we hope any algorithm will be applicable is so large that whatever sample we use, it must be considered too small. The only way to cope with this is to choose reasonably reliable evaluation principles. Many implementations of classification algorithms have a built-in function for a "leave-one-out"-evaluation (or "jack-knife" evaluation), which implies that from the data set $\mathcal{X}$ one instance is discarded and the rest used for training (or tuning), then it is checked whether the classifier makes the right decision for this instance. The instance is put back and another instance discarded and the rest used for training and the test is run for this other instance. This process is repeated $|\mathcal{X}|$ number of times. In this case, evaluation can be said to rely on a subdivision of the data set $\mathcal{X}$ into two subsets $\mathbf{X}$ (training data) and X (test data), where $|X| = 1$ and $|\mathbf{X}| = |\mathcal{X}| - 1$. However, according to e.g. Lavesson (2006, p. 24), "leave-one-out" is generally not considered reliable enough for algorithm independent performance estimation because it is often optimistically biased. It can however, be used for tuning, which is fortunate as it is easy to use.

For the estimation of performance, two other approaches are accurate. The first one is referred to as *10-fold cross-validation*. The idea of 10-fold cross-validation is that $\mathcal{C}$ is divided into 10 equally large and balanced subsets, 9 which are used for training and the tenth used for testing. Performing en evaluation ten times with each one of the subsets as testing data allows for average figures that are generally considered reliable, at least if the distributions of classes within the subsets are representative. The second approach is to choose an optimal subset $X \geq 30$ of $\mathcal{X}$, that is balanced with respect to the classes defined.

Second, we may compare the resulting figures with the result of other completely different approaches that are algorithmically simpler, which thus constitute *baselines*. The rationale behind this is of course that simpler models are always preferred over more complex models.

Unsupervised classification, i.e. clustering, is a completely different matter. It poses the problem that there is no inherent correct solution against which to measure the result, and consequently clustering usually relies more on cluster interpretation, rather than on performance measures. One could, of course, say that if the available data are annotated with the correct class, the data can be fed to a clustering algorithm without the class label, and evaluated against this label afterwards.

Clustering is otherwise commonly evaluated in terms of coherence and isolation measures. These measures are all based on feature values for the members of the clusters — measures that are often used by the algorithms themselves, which means that they are not algorithm independent (cf. Mirkin, 2005, pp. ff).

The question about which features are the most appropriate for representing genre adherence is still unresolved, and as long as it remains unresolved, clustering based only on feature-values is in theory bound to decrease accuracy to considerable degrees. However, clustering is a tempting method, because a) if it is succesful to a certain degree, the need for human intervention radically decreases, and b) previously unknown patterns may be discerned.

3.7 Concluding remarks on classification

The second and third research questions of this work relate to the task of classification in general. This chapter has sketched out how LIS approaches this as a human task and how classification is generally realized by means of algorithms.

The most striking difference between the library approach and the algorithmic approach is that the former stresses the design and human interpretation of classification schemes as such, while the latter is concerned mainly with the definition of features, and computational models which use these features for the task of classification. These two traditions are fairly distanced from one another, and it is difficult, for many reasons, to see where they can meet.

The next chapter gives an overview of how classification has generally been approached with respect to genre dimensions in its fairly recent computational research tradition.

Chapter 4

Previous research on the classification of documents along genre dimensions

In recent years, the interest in classification along genre dimensions and genre identification (as well as genre analysis) has increased considerably. At least since 1996, the *Hawaii International Conferences on System Sciences* have assigned a special track for "digital documents", where a subtrack has been "genre in digital documents". In 2007, several workshops (or similar events) have been held in conjunction with other academic events, such as the *ASIS&T*[1] *Annual Meeting* and *Recent Advances in Natural Language Processing*. Several theses have been written on the subject on both PhD and Master levels (e.g. Santini, 2007, Meyer zu Eissen, 2007, Boese, 2005, Rosso, 2005).

It has already been pointed out that research on algorithmic classification according to genre or the identification of genre differs in many respects.[2] This chapter will sketch out the ways in which the re-

[1] American Society for Information Science & Technology.

[2] The different wordings with respect to the topic to some degree reflect these differences. When genre is said to be identified, research tends to be less oriented towards browsing support.

search efforts differ. It will be structured in accordance with the three main issues for algorithmic classification of documents along genre dimensions: the definition of a set of classes, the choice of document features, and the choice of a classification model. In addition, as the sets of classes are dependent on the kind of document collection they are designed for, the kind and size of training and test data are important aspects with respect to these three issues.

Algorithmic approaches to classification along genre dimensions are, unfortunately, rare within LIS, even though user studies related to genre conceptualization have recently gained some attention (e.g. Rosso, 2005, Montesi & Navarrete, 2008).[3] Unfortunately, as with other kinds of user studies performed within LIS, they are seldom coupled with attempts to implement the results in real world applications. Implementations of algorithmic approaches to classification along genre dimensions are mainly to be found within the domains of computer science and computational linguistics. Here, for instance, Santini (2007) connects a small scale user study with experiments in classification along genre dimensions, and the domain of classification along genre dimensions does in fact need large scale user studies.

The foundation for algorithmic approaches to genre identification can be traced back to the traditions of empirical sociolinguistic research. The exploratory work by Biber (1988), based on factor analysis, is usually considered seminal. Factor analysis is a model that tries to find correlations between certain features from a set of *n*-dimensional data (i.e. *n* number of features characterizing one piece of coherent text), thus reducing the number of dimensions. Starting from a number of 67 features, Biber ended up with 7 groups of features (factors) that were used to characterize each of the 23 (spoken and written) genres in a collection of approximately 500, taken mainly from the Lancaster-Oslo-Bergen corpus of printed British English texts and the London-Lund corpus of spoken English (Biber, 1988, pp. 66ff). However, Biber's aim was, in a sense, the opposite of classification along genre dimensions, as he was mainly interested in characterising

[3] The study of Elsas & Efron (2004) could be considered an exception, but in this case the aim was to distinguish between web pages with contents of tabular, index-like or content-organised character — they do not claim to deal with genre.

language use, rather than identifying genre from language use.

It was not until the 1990s that similar methods were applied with claims to assist in information seeking tasks and improve information access. There are at least two obvious explanations for the fact that the concept of genre entered into information retrieval research not until some 40 years after the first experiments on topical information retrieval.[4] First, experiments must rely on sufficient computational resources in terms of large document collections and high processing power, which until the 1990s were insufficient. Second, as investigations into a research theme rely on their disciplinary acceptance and genre studies naturally embark from linguistics, the hostility towards research on language performance within linguistics as opposed to language competence, most prominently expressed by Noam Chomsky, has been disadvantageous (cf. McEnery & Wilson, 2001) up until fairly recently. The increased interest in corpus linguistics and availability of large corpora during the last decades have been important for the development of genre and register studies.

When linguistic attention has been drawn towards language performance in the 1990s, it quite natural follows that variation in language use becomes of greater interest. Results of this interest may be seen as providing computer scientists with sufficient background knowledge for one area of experimental application.

In trying to map algorithmic studies on genre discrimination we may note how four different experimental factors differ:

1. Genre space — The set of genre-based classes on which focus is placed.[5]

2. Feature set — The set of features derived from the documents to be used in classification, clustering or mere identification.

3. Document space — The kind and size of the corpus used for training and testing the performance of a classifier

[4] The origins of information retrieval research are commonly traced back to the Cranfield experiments in the early 1950s (Ellis, 1996, Chapter 1).

[5] These are sometimes referred to as *genre palettes* or *genre repertoires*.

4. Choice of computational model — The basic principles underlying the the training of a classifier

First, there is the question on how genre space is determined in terms of its nature and number of genres. Santini (2004b, p. 13,22) has pointed out there is no consensus in the way that genre space is defined, and that the number of genres studied is relatively low — it still is, at least compared to the amount of classes encountered in a library classification scheme.

Second, the features used in the identification process are more or less complex. In some cases simple lexical counts are used, and in some a plethora of features ranging from surface features, such as punctuation characters, to syntactical patterns are being used. In the latter cases the texts are preprocessed with e.g. parts-of-speech tagging or shallow parsing.

Third, the document spaces used differ much in extent and character, reflecting the nature of the genre space investigated. Not seldom, the document space is so small that it allows for almost no generalization. Document spaces are either carefully developed linguistic corpora or pulled together by the research teams themselves, often compiled from the web.

Fourth, the computational models used for the experiments differ, though they all may be described as more or less multivariate, reflecting the large number of features used. When some kind of classification model is used, it is polythetic and often more or less inductive. Supervised machine learning is the dominant approach for classification, which implies that the models are trained on a sample of preclassified documents before being applied on unknown documents.

An overview of how these factors vary in previous research on classification along genre dimensions is given in Table 4.1 on page 127.

4.1 Genre spaces

The understandings of genre, implied or explicitly declared, vary. The rhetorical and sociolinguistic understanding of genre as something extrinsic and related to the context of documents, i.e. reflecting communicative purpose, is apparent for e.g. Wolters & Kirsten (1999) and Rauber & Müller-Kögler (2001). Kessler et al. (1997), however, are careful to note that a genre (as applied in their experiments) requires that its functions must be "connected to some formal cues or commonalities" in order to count, and that the genre has to be extensible. In other words, a genre must, in their application, be algorithmically detectable and not restricted in the way that, for instance, the complete collection of Shakespeare's poems is. Sometimes, as for Bogdanov & Worring (2001) and Dewdney et al. (2001), a genre is defined by artefactual properties only. Folch et al. (2000) stresses this, as they do not talk in terms of genre at all, but in terms of text types. The term 'genre' is also sometimes substituted by 'style'. An interesting note to make here is that Karlgren (2000) observes that topic and style correlate to a certain degree, while Finn & Kushmerick (2003) hold that style is largely topic-independent — although they conclude that the performance of any genre based classifier is not topically independent.

The understanding of genre may at first be seen as practically irrelevant, as algorithms cannot in any way infer anything from what is not expressed in the document. Therefore it could be assumed that only form is what matters. But there are reasons to expect different realizations if the space of genres is understood as a stable space consistent over time and across disciplinary boundaries, or as a typified response to situations circumscribed by sociocultural expectations. In this case, clustering is possibly preferred to classification, as is the case for Rauber & Müller-Kögler (2001), who then use different colors and shifting nuances instead of descriptive labels to represent the different clusters, just because "for hardly any genre there is a strict and well-defined, non-overlapping set of criteria by which it can be described, making strict classification ... imposssible ..." (p. 2).

The genres investigated are usually rather few. Sometimes, experiments are performed from a binary perspective, weeding out documents that do not belong to the genre investigated. This was the case for Kaufer et al. (2005), who e.g. wanted to find technological reviews of personal digital assistants among documents that were topically restricted to that topic.[6] Lim et al. (2005), on the other hand, investigated 16 genres found on the web and Stubbe & Ringlstetter (2007) 32 genres. These studies are exceptional with respect to the number of genres investigated in one set of experiments.

In two recent approaches, Santini (2007) and Meyer zu Eissen & Stein (2004) both used the same corpus consisting of over 1200 documents precategorised into 8 web genres.[7] In addition, Santini compiled her own corpus of 1400 web documents, which were categorised into 7 perfectly balanced web genres.

Rehm (2002) only investigated the genre of the academic's personal home page, but introduces the notion of a "genre module" (inspired by the work of Haas and Grams (e.g. Haas & Grams, 2000)) that is somewhat similar to Görlach's bound text type. A genre module for Rehm is a type of genre-distinct part of a document, such as a list of publications, a navigational element, or the name of the author on a personal home page, with its own form and function. A document for Rehm is not necessarily confined to being equal to a computer file, but extended to include a whole web site. A genre module can then be regarded as something in between the document as a monolithic entity and its decomposition into functionally consistent parts, or even into text nodes, by means of markup.

In general, the genres investigated elsewhere reflect a rather simplifying approach to genre, where such culturally accepted terms as research reports, research articles, and home pages are taken to be fairly unproblematic designators of genres. The work of Kessler et al. (1997) expresses a fairly original approach. Instead of trying to find genres directly, they tried to detect what they called facets. A facet is a characterisation along one of three dimensions of a document. Texts may be either directed or broadcasted; they may be suasive or descriptive, and so on. Genres are then considered to be bundles of facets, as well as being characterized in terms of "brow" and narrative. An odd, and quite similar, approach is the one made by Finn & Kushmerick (2003), who use a collection of documents on politics,

[6] It must, however, be noted that the aims of Kaufer et al. (2005) were motivated by a quest for rhetorical knowledge and not for developing classifiers.

[7] The same corpus is used in the experiments of this work.

football, finance, and movie and restaurant reviews from the web, in order to classify documents according to whether they are positive or negative, or subjective or objective in their approach to the topics treated. It could be argued that while this is undisputably a detection of a non-topical aspect (and a classification of value for certain tasks), it is not really a matter of identifying genres. Their attempt becomes even more difficult when they train such a classifier on movie reviews and test it on restaurant reviews, which, it could be argued, are two different classes of texts within different genres. Not surprisingly, performance decreases considerably. It should be noted that the authors are not particularly clear about whether *reviews* belong to one genre or if *positive reviews* and *negative reviews* belongs to two genres. It is possible to interpret their wordings in both ways. However, it is clear that movie reviews and restaurant reviews are considered topically different but not of two different genres.

There is a strong tendency to use either precompiled linguistic corpora containing balanced samples of everyday language use (Karlgren, 2000, Kessler et al., 1997, Wolters & Kirsten, 1999, Wastholm et al., 2005) or newspaper material (Ihlström & Åkesson, 2004, Rauber & Müller-Kögler, 2001, Argamon et al., 1998, Mehler et al., 2007), such as from the Wall Street Journal (Stamatatos et al., 2000). As a result, the choice of genre space is highly dependent on the nature of these corpora, for which precategorizations are often performed with aims other than the classification along genre dimensions.

4.2 Document spaces

Another varying factor of the experiments is thus the size and nature of the collection of documents used in developing and evaluating the algorithms. A collection of documents needs to be fairly balanced with respect to the distribution and number of target genres investigated. Furthermore, it has to be precategorized and each document labelled with a class designator, at least if supervised classification is applied. One may discern two tendencies here. Researchers either use readily available corpora precompiled for linguistic studies or compile and annotate new corpora themselves. In the first case, the classification along genre dimensions becomes highly influenced by categorizations used for other purposes, which do not always fully comply with the genre notion applied. Karlgren (2000), in his initial experiments, performed as early as 1994 together with Cutting, used a sample of approximately 500 documents from the Brown corpus, as did Kessler et al. (1997). The Brown corpus demonstrates a broad categorization which partly reflects topical subdivisions, rather than genre subdivisions. Wastholm et al. (2005) used the Swedish counterpart to the Brown corpus (i.e. SUC), and Wolters & Kirsten (1999) a German counterpart, both of similar size and nature.

When linguistic corpora are not used, collections may be adjusted more easily to the purposes of the classification along genre dimensions, in order to facilitate validity and exportability. The main problem with individual compilations lies in risking more or less unreliable precategorizations, due to the inherent vagueness of individual genre conceptualizations. This is why several parallel annotators may be used. Meyer zu Eissen & Stein (2004) did try to compensate for this by having at least three annotators. However, as will be shown in Section 5.3, even this is sometimes not enough.

The size of the collections is another crucial issue, because both high-dimensional document representations and an increased number of genres increase size requirements. This sparse data problem is especially crucial for certain parametric classification methods, such as Bayesian classification. In addition, especially as web genres are concerned, genres are constantly changing, which requires that the size of the data covers enough individual variation. A partial reason for the lack of large scale corpora designated for genre studies is probably the amount of work that has to be invested. Sampling and annotation takes a lot of time.

The standard linguistic corpora may be said to reflect a certain restricted area of documentation practices for which the common size of 500 samples may or may not be enough. It depends on which claims are being made. The even smaller samples used by e.g. Bogdanov & Worring (2001), Stamatatos et al. (2000), and Kaufer et al. (2005) do not seem large enough to allow for generalizations. However, their scopes are fairly limited and they do not make large claims.

Dewdney et al. (2001) used a sample of 9705 documents from the general categories of advertisements, bulletin board messages, FAQs, message boards, radio news, Reuters newswire material, and television news. Lee & Myaeng (2002) used two samples of approximately 7000 documents in Korean and English respectively, collected from the web. Mehler (2007) used a newspaper collection of over 32000 texts. These are exceptional with respect to the size of the collections.

It is highly unfortunate that there have not been any available corpora that allow for benchmarking with respect to genre classification, like the TREC collections do for information retrieval research. However, fairly recently, Santini (2006, among others) has recognized this problem and now offers a few preclassified collections on her web site, compiled by different research groups.

4.3 Features

The number of features used in genre identification is generally high, because a class of genre artefacts is usually considered not to be reliably identified by just a few properties, such as a restricted number of keywords. The number of features used by Biber in the 1980s was 67, but recent efforts tend to regard even more features. For instance, Finn & Kushmerick (2003) used 152 features consisting of parts-of-speech categories, function words, punctuation characters and "document-level statistics". Rauber & Müller-Kögler (2001) used around 200 features, mainly consisting of "text complexity measures", counts of special characters or punctuation signs, certain genre-specific keywords (such as § for legal texts), and specific technology dependent markup. However, the number of features used is dependent on algorithms, and on whether e.g. features represent one-token lexicals or groups of similar lexicals. The polythetic nature of classification along genre dimensions seems to necessitate a high dimensionality, which can however be reduced by different techniques, such as *Factor Analysis*, *Singular Value Decomposition*, or *Principal Components Analysis*. Most research must be said to rely on degrees of correlation between features that may have a fairly skewed and unpredictiable distribution over large samples of data. In fact, most importantly, Kim & Ross (2008) observe that the optimal choice of features depends on the genre space.

Stamatatos et al. (2000) illustrate what may be one of the crudest attempts to the choice of features. They used word counts and investigated how certain words were discriminative for certain genres. Words that within a group of documents with the same topic demonstrate a high frequency within one genre, but a very low frequency in other genres, were considered discriminative. Their result is then a set of 30 discriminative words. To this kind of features they added figures on the occurrence of punctuation characters. Not surprisingly, this poses an overfitting problem, i.e. the performance of the classification algorithm becomes too much dependent on the training material. The same result is observed by Kessler et al. (1997), but it should be kept in mind that both Stamatatos et al. and Kessler et al. worked with small samples. Otherwise, Kessler et al. (1997), in addition to words and punctuation counts used what they refer to as derivative cues, which is a combination of lexical cues and character-level cues — sometimes similar to traditional text complexity measures, such as the proportion of long words. A fourth group of features, referred to as structural cues, is used by some researchers. Karlgren (2000), Dewdney et al. (2001), Argamon et al. (1998), Wastholm et al. (2005) all used part-of-speech tags in their experiments. A drawback for real world applications is that this kind of features requires part-of-speech tagging or parsing as a computationally expensive preprocessing. A fifth group of features used by, for instance, Santini (2005), Lim et al. (2005), Elsas & Efron (2004) and Rauber & Müller-Kögler (2001), are counts of certain HTML tags. The greatest interest in these cases is on markup for hyperlinking and embedded images. Lim et al. (2005) assigns particular interest to the URLs of the hyperlinking tags and take account of whether they refer to documents within the same domain or not. Mehler et al. (2007) have, most notably, explored what they refer to as the logical document structure as a basis for classification, where these features are derived from e.g. markup. Unfortunately, they do not describe how the source documents are authored with respect to markup performance. In addition, the training and test data are derived from the same source, the Süddeutche Zeitung, which can be considered very much constrained by a single editorial policy.

Besides these five groups of features, there are some scholars who focus on the appearance of documents, and nothing else. Ihlström & Åkesson (2004) do this, but it should be remembered that they do not primarily aim at classification for information seeking tasks. Bogdanov & Worring (2001) use Random Graphs to model the physical appearance of documents. Power & Scott (1999), Hu et al. (1999) are two other examples. Kim & Ross (2007) use a combination of textual surface features[8] and image features.

A final uncommon and interesting kind of features, that is derived from the speech acts conveyed by certain lexical indicators, is used by Goldstein & Evans Sabin (2006) in the categorization of email messages.

The results of all this research give no significant clues to whether some set of features are better than others, but most research reports on that combinations of different types of features yield better results in classification tasks.

4.4 Models

The models from which algorithms are elaborated vary a lot, but they are all in general counted among the machine learning techniques. K-means clustering (Santini, 2005), EM clustering (Kim & Ross, 2007), and Self-Organizing Maps (Rauber & Müller-Kögler, 2001) are the clustering techniques applied. These unsupervised models are interesting as they do not rely on preclassified training data, and might thus comply with the vagueness of genres. However, they are sometimes used for preliminary investigations of experimental data, in order to explore any inherent structure of the data. Santini (2007) used K-means to explore how genres were correlated to the Werlichian text types.

[8] The term 'textual surface features' denotes the absence of linguistic categorizations.

Otherwise, Discriminant Analysis (Karlgren, 2000), Principal Component Analysis (PCA), or Factor Analysis (Biber, 1988) are often used in order to find discriminative features, besides being used for dimensionality reduction (e.g. by Santini, 2007), whereas there seems to have been a preference for Decision Trees (Argamon et al., 1998, Dewdney et al., 2001) and Naïve Bayes (Dewdney et al., 2001, Lee & Myaeng, 2002, Wastholm et al., 2005) in the actual classification tasks. In recent times, Support Vector Machines (SVM) have become popular (Meyer zu Eissen & Stein, 2004, Santini, 2007, Mehler et al., 2007, Goldstein & Evans Sabin, 2006), as it is generally considered one of the most successful model for classification tasks in general. These latter examples are all counted among the supervised machine learning techniques.

When choosing among the models from which algorithms may be derived, there is no clear indication that one model performs better than others. Even though there are some examples (e.g. Dewdney et al., 2001, Lim et al., 2005, Kim & Ross, 2007) where several models have been tried on the same material, and in that case one of the models performs significantly better, it seems that the results are highly dependent on a combination of the character of the document space, the genre space, and the features.

Table 4.1 summarizes how four experimental factors differ within a sample of previous research on the classification along genre dimensions. The figures in the table only indicate the variation. For instance, many projects have experimented with different dimensionalities of the feature space or different data sets. The numbers given are only approximate values, and the table should not be considered exhaustive in any way.

4.5 Concluding remarks on previous research on the classification along genre dimensions

Previous research on the classification along genre dimensions indicates that there is not much we can know for certain about which classification models perform best or what kind of feature sets are the most

	$\lvert\mathcal{C}\rvert$†	$\lvert\mathbf{x}\rvert$††	$\lvert\mathcal{D}\rvert$†††	Model
Karlgren (2000)	2/4/16	40	500	Discriminant analysis
Kessler et al. (1997)		55	500	Logistic Regression / Neural Network
Argamon et al. (1998)	2	>1000	400	Decision trees
Wolters & Kirsten (1999)	9	>50	500	MBL (several variants)
Stamatatos et al. (2000)	4	<60	160	Discriminant analysis
Rauber & Müller-Kögler (2001)		<200	1000	SOM
Dewdney et al. (2001)			10000	Decision trees, Naïve Bayes
Lee & Myaeng (2002)	7		7000	Naïve Bayes
Finn & Kushmerick (2003)	2	152	800 & 1300	Decision trees
Meyer zu Eissen & Stein (2004)	8	35	1200	Neural Network / SVM
Wastholm et al. (2005)	9		500	Naïve Bayes
Shepherd et al. (2004)	4(3)		>300	Neural Network
Boese (2005)	10	78	>300	Logistic Regression
Lim et al. (2005)	16	>300	1200	k-NN
Kim & Ross (2007)	2		750	Naïve Bayes, SVM, Random Forest
Santini (2007)	4-8	>200	>1000	K-means, SVM

Table 4.1: Overview of variation in previous genre research. $|\mathcal{C}|$ is the number of genres investigated, $|\mathbf{x}|$ is the dimensionality of the feature space, and $|\mathcal{D}|$ is the size of the corpus.

† When figures are missing, numbers are not applicable.

†† When figures are missing, numbers are not reported or not relevant due to the character of the implementation.

††† Figures are highly approximative.

efficient. What we do have indications on, however, and this is what this work leans on, is that

- since classification along genre dimensions is much in its infancy, any research must account for the fact that when the resulting set of classes is increased, task complexity also seems to increase and it is therefore wise to keep this cardinality within reasonable limits;
- there is no clear evidence that there is a set of features (or a set of feature types) that generally performs best — neither have all possible kinds of features been fully explored;
- the kind of target genre space to which classification is to be applied highly affects the performance, and different genre spaces seem to need different feature sets.

The first observation affects the choice of experimental setup in the way that the notion of genre is somewhat simplified and the otherwise more reasonable high granularity is refuted. This is motivated by the second observation, that novel features are worthy of experimental investigation. The third observation motivates investigations of varying genre spaces rather than focusing on one particular genre space.

Part II

Experiments in Document Genre Classification

Chapter 5

Experimental setup

The second part of this work will start by sketching out the framework for a set of experiments intended to facilitate the answers to the fourth research question: **How do different definitions of genre spaces, classification models and document features influence document genre classification?**

5.1 Theoretical premises

In the first part of this work, a theroretical investigation and a review of previous research on classification along genre dimensions have been undertaken in order to establish a firm theoretical and multidisciplinary framework for the concept of genre in the context of classification. An important question is whether any theoretical conclusions can be drawn from this that must be taken into account when designing an experimental framework.

Attempts have been made to show that the concept of genre deserves attention as denoting an abstract phenomenon that arises from the conflation of several configurations around social activities which make use of documents, and that this concept must be distinguished from the types of documents involved. However, this distinction is often neglected in library practices and in research on genre classification, so that the concept of genre is actually identified with a class of documents. This is no coincidence, since the documents of most genres are expected to be highly typified with respect to their linguistic, paralinguistic, and technological appearences. The documents of a certain genre are recognized by their appearances, their form, a fact which is made use of in applications of the classification along genre dimensions.

The crucial question, given the understanding of genre as social action, is whether artefactual typification with respect to lingustic, paralinguistic and technological expressions is sufficient in order for an algorithm to arrive at reasonably correct mappings of documents to genres. It seems evident, judging from the theories and the research reviewed in the previous chapters, that one needs to be extremely careful with the choice of target genres and feature sets, and to explore different combinations of genres and feature sets.

This is actually the core challenge of classification along genre dimensions, similar to the challenge of classification along topical dimensions, where word and phrase distributions are usually the cues for finding out what a text is about.

Another problem has arisen as a result of the explorations of the concept of genre in previous chapters. This problem concerns the fact that any genre needs to be justified with respect to human expectations and the aims for which humans search for documents, if it should be of any use in information seeking contexts. The most common way of performing this task is to proceed from what one percieves is the name of a target genre. However, as has been discussed, names are often confusing, such as in the case of the label "bibliography". The purpose and the target users of a bibliography in a thesis are quite different from when a national bibliography or a subject guide is concerned, even though they may very well be sought for by means of the label "bibliography", and express very similar appearances.

Subsequently, if one needs to define a set of target genres for experimental research, it seems important to take care and define and describe the context in an elaborated way, rather than only giving the name or form of a class of documents, so as not to create confusion.

For the experiments of classification along genre dimensions pursued in the next part of this work, the main difficulty and the problem referred to above are tackled by means of putting stress on the variation of genre granularities and different feature sets, while being careful with identifying a contextual configuration, when defining genres.

The state-of-the-art classification models pose another difficulty that is treated by means of a simplification. Even though it must be admitted that a document may appear in different contexts and subsequently adhere to several genres, existing implementations do not allow such an approach without extensive modifications to the source models. Multi-genre documents are thus sought to be avoided as far as possible.

5.2 Defining the setup

The fourth research question in itself calls for an approach in which different genre spaces, classification models and feature sets are set to vary in different combinations. The most common available models for classification have been briefly presented in Chapter 3; and in Chapter 4 the common models, types of feature sets, and genre spaces investigated in previous research on classification along genre dimensions have been reviewed. Given the models, genre sapces and features presented in Chapter 4, the number of possible combinations is high and has to be restricted in some way.

The experiments of this work are highly influenced by an intention to gain extended knowledge on the implications for classification of, on one hand, the view on the concept of genre as an expression of social action and, on the other hand, the relatively unexplored notion of document structure in relation to genre, presented in Chapter 2. Thereby the variation of genre spaces and feature sets becomes the main focus for the experiments. It is deemed more interesting to study the variation of target genres and feature sets, than the variation of classification models, even though it must be remembered that some models may be more or less biased towards managing certain genres or feature sets.

The general research question additionally calls for an approach that cannot avoid the study of classification results by means of performance estimators, such as those presented in Chapter 3. Otherwise there would be no way of quantifying the influence of different definitions. This is the means whereby any tendencies of the influence of different definitions may be determined. It is , however, somewhat unfortunate that such estimators require that the operational understanding of genre adherence must be allowed a somewhat undue simplification, so that the mapping of one document to one genre is not fuzzy or undetermined, but either true or false. For instance, a document that paraphrases the artefactual type of another genre cannot but be assigned to the genre which it paraphrases. Additionally, given a low genre granularity, the documents assigned to such a genre class will naturally be heterogeneous with respect to form, purpose and target community. This is the case for the class of "articles" in one of the precompiled corpora of the experiments.

Normally one could expect that a set of experimental questions should be stated before the setup of the experimental framework. However, the questions that can be formulated and the ways in which the framework might be set up are dependent on one another, and restricted by practical matters, such as available algorithm implementations and data. This is why this chapter is concluded, rather than started, by the definition of two experimental questions.

The general research question reflects what has been pointed out in the previous part of this study, namely that there is a set of main issues in setting up a framework for experiments in classification:

- Choosing or designing a model for classification ;
- Choosing or compiling a document collection or corpus, i.e. a set of documents for classification ;
- Deciding on a suitable feature set for input to the classification model.

In addition, if the data set is a corpus, the set of classes to be investigated are given by definition and the correct class assignments known beforehand. Otherwise, one also has to

- Decide upon one or more sets of classes (genre spaces) to which the documents of the collection are to be mapped.

The choice of a corpus is motivated below in Section 5.3. The classification models applied are shortly described and motivated in short in Section 5.4. The feature derivation is presented in Section 5.5.

5.3 The data set

The choice of experimental data for the study of classification along genre dimensions is far from straighforward, for several reasons. Reliability issues require e.g. that others should be able to reproduce the experiments and compare the results, so the data must be readily available. Validation issues require that the data are appropriate for the research questions posed.

The criteria for reliability points towards choosing an existing corpus. If there are collections of documents that have been compiled and annotated by others, the gain is threefold.

1. The great amount of time it takes to annotate each document in advance is saved.

2. The consistency, intersubjectivity, and accuracy of the annotation can be improved.

3. Benchmarking is made possible.

In choosing an existing corpus, it is relevant to estimate the value of its predefined genre space with respect to the experimental research questions. No such guidance is available as yet, and the choice thus becomes a matter of being pragmatic towards an ideal situation. The choice actually made here is related to a common trait for academic libraries, namely to support students and researchers in their work. If this is the focus, the matter of trustworthiness is also in focus. If a document is to be supportive in studies and research, it has to be trusted with respect to a set of criteria. This is is no place to elaborate on the tricky concept of trustworthiness, but Francke (2008, p. 115ff) gives a thorough account of its importance with respect to open access journals and the notion of cognitive authority, borrowed from Patrick Wilson.

What can be considered useful material for a researcher or a student is of greatest interest, as these activities call for a high grade of consideration of the genre of the document. For instance, private portrayals, i.e. personal home pages, can be valuable in an information seeking task as pointers to possible ways of thinking or getting informed.[1] However, such material is not commonly used in scholarly writing to support an argument, and is thus somewhat out of the scope of the focus of this work.

Therefore, when looking for existing corpora, it is relevant to ascertain that they contain a certain amount of scholarly material of the kind that can be used in e.g. writing papers, and that the corpus in question includes a genre in which such material is a common trait. It is likely that e.g. the partitioning of a collection of FAQ pages (Frequently Asked Questions) is less valuable than the partitioning of a collection of scholarly writings. Additionally, in the user study of Meyer zu Eissen & Stein (2004) it was observed that what they referred to as classes of “articles” and “scholar material” were considered by users as being more “favored genre classes” than e.g. ”link collections” or “news” material.[2] It is assumed that for FAQs, web sites with discussion postings, and web sites that offer shopping possibilities, the contents are more likely to be accessed on topical search criteria only, and further divisions are thus less valuable than if a user is confronted with a mixture of reviews, research articles, scholarly introductory material to a specific topic, editorials, project reports, technical briefs etc. — all of which can also be considered broadly uniform with respect to communicative situation. These fine-grained classes of documents seem to constitute interesting and valuable clusters for this work, if it is possible to arrive at such clusters.

Some validity constraints, concerning the task of delimitation, should be added to these guidelines. Ideally, in order not to have individual instances contributing too much or too little to the result, the

[1] ’Personal portrayal’ is in fact a significant label for a genre, while ’personal home page’ is rather a label of a class of documents.

[2] It is , however, not completely clear if these favored classes were in themselves considered useful or if these classes were of most interest for genre based classification.

data set needs to be balanced, that is, the different strata (i.e. genres or classes of documents) need to be fairly evenly distributed over the complete sample. In other words, the entropy needs to be maximized. Even though in explorative approaches the data are often already given and the objectives may be e.g. only to find a useful way of representing the structure of the data or to describe the character of clusters (cf. Mirkin, 2005, p. 3), a good guess on an even distribution of different potential clusters is fortunate. Another validity aspect is that both classification and clustering set criteria for the size of the data set. It seems important that the number of instances in the dataset is at least greater than the number of features used, in order to avoid overfitting the algorithm to an empirical bias (cf. Alpaydin, 2004, p. 144).

At the time when this work began, early in 2005, there was only one corpus available that met the critera set out above, namely the *KI-04* corpus. Boese & Howe (2005) give an overview of existing corpora at this time, where it is shown that either the corpora are too small in regards to the number of documents (with respect to the number of classes contained), too old, or simply unavailable. Since that time, several novel corpora have been compiled, e.g. the 7-web corpus by Santini in 2005, but none of these fully comply with the criteria for this work, as of 2007. A directory of available corpora is maintained in the WebGenreWiki (2008). An interesting initiative is the fine-grained corpus compiled by Stubbe & Ringlstetter (2007), which for the moment, though, is too small given the amount of genres covered.[3]

As the significance of different feature sets is part of the objectives and features rely in large on the use of natural language, there is one more aspect to consider in the choice of data. The use of language varies on many levels. Several of these levels depend highly on which natural language is used. The lexical repertoire within e.g. English and German is, of course, different. This is also the case for the use of punctuation, to some extent. If features are derived from natural language use, their application is clearly not language-independent. Subsequently, given that the size of the document collection has to be restricted, it is probably wise to refrain from having different natural

[3] 32 classes of 40 instances each.

languages represented. In addition, feature extraction relies on the extraction of character data, and non-English languages often demonstrate problems with regards to character respresentation, which is why English seems the simplest choice here.

Finally, the file format is also of importance, from two aspects. First, as one of the experimental questions is related to features derived from the (logical) document structure, it is necessary to employ some kind of declarative markup, from which the features can be derived. Flat text files, pdf documents, and WORD files all need thorough preprocessing in order to convey such features, besides the fact that pdf documents and WORD files need to be converted into files with a standard character encoding. Second, the majority of documents that need genre classification are encoded in some version of HTML.[4]

5.3.1 The *KI-04* corpus

Taking account of what has been recapitulated above, a precompiled collection of documents, the *KI-04* corpus, has been chosen. In the beginning of 2004, a collection of 1295 documents were compiled by a German research group in order to be used for experiments in genre classification. Their work was reported by Meyer zu Eissen & Stein (2004). The compilation was accomplished by collecting bookmarks from five individuals and somewhat extending the bookmark collections, in order to get a balanced collection according to eight coarse-grained categories of documents. These were labeled, 1) articles, 2) download pages, 3) link collections, 4) private portrayals (i.e. approx. “personal home pages”), 5) pages with group discussions, 6) help pages, 7) non-private portrayals, and 8) shopping pages. Note the use of ’pages’ here, because it is a matter of individual HTML files. Instances are individual files rather than coherent web sites. The group also consistently refers to the categories as “genres”, not classes of pages belonging to a genre. These eight “genres” will mostly be re-

[4] It may be important to point out that a corpus of HTML documents using the same XML version of HTML would be ideal, but very hard to accomplish. Less than 15% of the material from open access journals in the study of Francke (2005) are valid XHTML documents.

ferred to as the *KI-04 classes*.

Santini (2007), who used the same corpus for her PhD thesis, found that a subset of these documents was void of content and ended up with 1205 documents. Antivirus software found that two of these remaining 1205 documents were infected and they were subsequently discarded before the corpus was used in the experiments of this work. The distribution of these 1203 instances over the eight categories is given in Table 5.1. Its skewedness is not optimal, but in order to respect the integrity of the data, there has been no attempt to balance the corpus.

name	**# of instances**
articles	127
download pages	151
link collections	203
private portrayals	126
pages with group discussions	127
help pages	139
non-private portrayals	163
shopping pages	167

Table 5.1: Class distribution in the *KI-04* data set

The 1203 documents have been preclassified and annotated by the compilers. That is, each document was labeled according to which of the 8 classes it was assigned to, its textual contents stripped from markup and added as a kind of comment to a file header (see Figure 5.1).

Meyer zu Eissen & Stein (2004) and Santini (2006), who both performed experimental studies with the *KI-04* collection, attained maximum accuracy figures around 70%, which is low compared to other previous experiments in genre classification. A second corpus used in Santini's experiments, with the same feature sets, resulted in much better performance figures. Santini (2006, p. 164) attributes this to the compilation of the *KI-04* corpus, that it has to do with a decreased objectivity in annotation and a non-consistent genre granularity.

A closer look at one class reveals the heterogeneity of the class

```
<!-- <DOCUMENT>
        <FILE>
                /home/smze/tmp/roman-source/corpus/0654181015.html
        </FILE>
        <URL>
                http://www.geocities.com/algnotes/whatuse.html
        </URL>
        <TITLE>
                Use of Algebraic Geometry
        </TITLE>
        <GENRE>
                articles
        </GENRE>
        <PLAINTEXT>
 Use of Algebraic Geometry /... /
        </PLAINTEXT>
        <CONTENT>
-->
<HTML>
/ ... /
```

Figure 5.1: Sample header from the *KI-04* corpus

annotation. Within the *articles* class there are samples that show unambiguous marks of being e.g., glossaries, fictional excerpts, tables of contents, and manuals for a programming language or a piece of software. It seems that these kinds of text are all preferrably assigned to the *articles* class, as no other class is more appropriate — they do not meet with the definition of articles given by the compilers themselves: “Documents with longer passages of text, such as research articles, reviews, technical reports, or book chapters” (Meyer zu Eissen & Stein, 2004, p. 6). Additionally, in the *articles* class there are also one or more instances that can easily be regarded as instances of other classes defined for the corpus, such as *help pages*, *private portrayals*, *non-private portrayals*, *link collections*, or *discussion pages*.

Furthermore, the collection contains a few documents written in German. There are, for instance, three documents in German in the *article* class. For validity reasons, these are discarded in the experiments that focus on *articles*, while they are retained in the experiments with the complete data set, only for the sake of complying with previous research.

For the purposes declared in Chapter 6, the *KI-04* is eventually extended by 54 other documents added to the *KI-04 article* class. These have been collected on the basis of a stratified search session in Google, where a) a set of HTML research articles was downloaded from several quite arbitrarily chosen open access electronic journals (no more than one was chosen from each journal), b) a set of "project descriptions", "introductory" material, "technical reports", and "technical briefs" was identified and downloaded on the basis of being first encountered using these labels as queries provided to Google, with the option of excluding pdf files enabled, as only HTML documents were of interest. These were annotated by the author according to their labels.[5]

5.3.2 Corpus reannotation

The annotation of *KI-04* uses a loose undocumented SGML-based annotation embedded in the documents in a comment section (see Fig. 5.1). This is unfortunate when it comes to reannotation. There are several reasons for this, and the main problems with embedded annotations concern readability and processability, as pointed out by e.g. McEnery & Wilson (2001, p. 38).

- Embedded annotations require that the documents are being processed for reannotation, which may be done by means of a scripting procedure or by means of opening each document and editing its content. Both processes require much human time and effort.

- Embedded annotations make the task of extracting data from the documents equally time-consuming, especially since the documents themselves do not conform to predictable markup standards that would otherwise have made it easier.

- The alternative, stand-off annotation, makes parallel annotations highly plausible.

For these reasons, a simplified document type definition (a DTD, see Chapter 2.3 for this concept) has been written in order to store the annotation of the documents (i.e. their class assignments and any remarks on these assignment) in a separate file, and the class definitions

[5] Mostly, these labels occurred as HTML titles.

in another separate file. The DTD is given in Appendix A together with some exemplifying snippets from the XML files.[6]

The 8 classes of the *KI-04* corpus do not appear to have been explicitly and clearly defined to the extent that a consistent annotation could be applied. What is to be understood with a *private* and *non-private portrayal* may seem evident, as well as what is implied with the genre labels *discussion* and *help pages*, but obvious differences may be discerned between the documents in several of these classes. The class of discussion pages can be subdivided in three large groups: 1) documents containing the contents of only one posting, 2) document containing a set of individual postings, commonly on the same subject and several being responses to a preceding posting, and 3) indices, or listings, of pointers to individual postings or groups of postings. For an information seeking task it is perfectly possible to imagine a user who is content with any one of these types. However, strictly speaking, indices to postings and the postings themselves are two different phenomena with respect to their functions, even though they may be equally relevant for some tasks. The problem is not that they should be considered different, but that their derived features are most likely to differ considerably.

The class of articles is a different matter. An announcement for a book on computational intelligence (article_4782928506) cannot easily be seen as an artefact adhering to the same genre as a short instructive introduction on how to calculate percentages and probabilities (article_1645868862).[7] The purpose of the first document above is clearly to promote buying of the book, and the purpose of the second is educational. Even though this could have been seen as a misclassification, none of the documents would have been better assigned to any of the other 7 classes.

There are even more instances in the article class that illustrate the same dilemma, i.e. that are equally unfit with each one of the classes in the corpus.

[6] A set of XSL stylesheets for different kinds of transformations of the contents of the two files has also been written and may be requested from the author. This set includes a stylesheet for transformation into a simplified XML TOPIC MAP.

[7] The references given in parentheses are identifiers for documents in the corpus.

In order for the experiments to be valid, i.e. to illustrate what they ought to illustrate, documents that do not reasonably fit into the classes have to be removed, so that they do not play inadequate roles of being outliers. Other documents have to be reclassified in a less coarse-grained fashion.

The complete article class, originally containing 127 documents, has been thoroughly examined and given one, and only one, new descriptive label, derived by the author in an attempt to arrive hermeneutically at a fine-grained genre space that seems reasonably intuitive, given the understanding of genre applied in this work. In this reannotation process, no respect to balance has been given.

The result of the reannotation was a set of 30 genre labels, distributed over 127 documents. It is obvious that dividing a set of 127 documents into 30 classes by means of algorithmic approaches is unrealistic, especially as in several cases only one or two documents have been assigned to a class.

It has been proposed that a genre can be considered a supergenre or subgenre with respect to another genre. It could be assumed that in this case each of the 30 genres, except for the obvious anomalies or misclassifications, can be considered a subgenre of the article genre. This is not argued by implication here — super- or subordination is not considered. They fall into the *articles* class just because they are set apart from the rest of the corpus by the original compilers, and are treated as such only for reasons of comparison.

However, the refutal of the existence of supergenres and subgenres does not imply that there is no way of grouping sets of genres on the basis of other aspects on language use and documentation strategies. For instance, some genres can be characterised as being realized in artefacts that employ predominantly the argumentative text type (according to the notion of text type, recapitulated in Chapter 2). Many contributions to electronic journals are predominantly argumentative. According to the theory of facetted classification (briefly recapitulated in Section 3.4), dominant text type can be considered a facet, as well as a genre (cf. how Santini, 2007, regards text type as an intermediary level between genre and features in Section 11.2 of her thesis).

For the reason of making classification realistic, the 30 genre classes are grouped in four overarching classes based on a slight reconsideration and pragmatic modification of the notion of a text type. An enumeration of these 30 genres, together with their number of occurrences, are given in the long table starting at page 147.

1. The group of mainly *instructive* documents that have a clear educational context or introductory character with respect to a concept, phenomenon, or directions for how to do something. Among the examples are text book material and manuals.

2. The group of mainly *argumentative* documents that are often focused on supporting an argument. Among the examples are the paper genre and the thesis introduction.

3. The group of *reporting* documents that are predominantly narrative and/or descriptive in their way of reporting what has been done. Among the examples are the technical report, the project description, and the technical brief.

4. The group of nonconsistent *unfit* documents that do not naturally fit with what is commonly implied with the word 'article'. Among the examples are the table of contents, the bibliography, and the dictionary.

First of all, there is a set of classes that are highly instructive and explanatory and aimed at helping the reader in some way, either with a practical matter or with understanding a conceptual issue.

There are around twenty documents that are of a **text book** character. These documents are highly instructive and aimed at an educational context.

Two documents are deemed to be equally aimed at an educational context with an instructive purpose, even though they are produced to support a lecture. The difference from the text book class is that the documents of this **lecture notes** class do not function without the lecture.

Sixteen documents are very similar to those of the text book class, but these **introductory samples** seem to be much less tied to an educational context and stand alone as online resources for whoever needs an explanation of some concept or phenomenon.

The seven **how-to**s are similar to the introductory samples, but focused on doing rather than explaining, mostly with respect to the management or installation of software or a programming language.

Three documents are part of a **manual** class, and are a kind of **how-to**s, but more encompassing in their situational scope.

Two documents each have been assigned to the **FAQ** and **discussion** classes. These are all *responses* to questions posed. The difference between them is that documents of the latter class contain answers from several individuals in which new question may also be embedded, whereas in the former there is only one individual who answers a question.

These classes may all, except for the last two, be considered part of a more generic class. A problem is, however, that there are many documents that are only fragments of a larger whole.

Four classes fall into a category which may be considered as typical for what is probably perceived as typical articles in that they are highly argumentative, promoting a certain idea or argument.

The **paper** class consists of documents that are prepared for inclusion in a journal and follows, in general, the scientific criteria for the subject domain which they are aimed at.

In the **argument** class, document authoring seems to lack any obvious purpose of having the text included in a journal. The author argues for an idea, but only seems to intend the document for open publication on his or her own website.

There is also one class, **thesis intro**, which has only one member, an introduction to a thesis, declaring the motivation for the problem stated in the thesis.

In addition, **editorials** are a common trait in journals and usually introduces the contents of a journal issue by means of enumerating and commenting on each contribution, but may also include a more argumentative part.

Four classes fall into a category of enumerating documents whose aim is to guide the reader to other sources and not intended for sequential reading.

The **bibliography** class contains ordinary bibliographies, but also link lists. The purpose is clearly to inform the reader of what could be of interest. The context is, in general, a constrained subject domain.

The **table of contents** class contains documents that perform the same function as a bibliography, even though pointers are restricted to sections of one document.

The **book announcement**, on the other hand, is a bibliographic item with an extended description. Its purpose is, compared to the bibliography, more or less commercial, promoting the buying of a book.

A **link list** differs from a bibliography in that it contains almost no description of to what its items point.

Four classes are highly narrative or expository, thus the documents have more of a reporting character.

The **project description** describes the progress and aims of a project. It is reporting and the aim is very similar to a non-private organizational home page.

A **technical brief** is quite a bit like a project description in terms of purposes, but it is generally the product that is described, not the project work or the working group.

The **technical reports** class includes e.g. technical specifications, and the documents are sometimes more instructive than the project descriptions and technical briefs. They could be expected to be approximately conformant to the requirements of the Technical report standard (American National Standards Institute, 2005), but this is not a requirement for inclusion.

Several documents do not fit well within any of the classes above.

Two documents are only **narrative pieces** and lack any obvious purpose at all, four documents are **glossaries** or dictionaries enumerating terms and their definitions and explanations, two documents are **applications** that receive input from the reader and return different results based on that input, and two documents would be better classified as **non-private portrayals** of a commercial character. There are also a **letter to the editor** of a newspaper, a mathematics **quiz**, a **redirect** page, a **private portrayal** about the research of an individual, and two documents that describe some data which are not present, just like a figure or table **caption** does.

Even apart from the obvious misclassifications, the heterogeneity of the article class is far from satisfactory when it comes to training an algorithm to classify documents according to genres.

In the experiments, the article class is significantly reduced. The instructive classes, the argumentative classes, and the descriptive/narrative classes are the only ones retained, which leaves us with only 91 remaining documents in the article class. Among these, two are authored in German, and excluded as there are several features used that are dependent on the lexical repertoire of the English language. Thus the number of remaining documents is 89. The obvious misclassifications could have been reallocated to their correct class, but since a balanced corpus is sought, there is no point in doing so, given the comparatively large size of the other *KI-04* classes.

Label	**Instructive**	**Argumentative**	**Descriptive**	**Unfit**	**#**
Paper		x			19
Textbook	x				18
Introductory	x				16
Argument		x			8
Manual	x				3
How-to	x				7
Project description			x		7
Technical brief			x		7
Editorial		x			3

Continued on the next page

Continued from the previous page

Letter to Editor		x			1
Technical report			x		2
Thesis introduction		x			1
Press release				x	2
Abstract				x	1
Q/A				x	2
Table of contents					3
Link list				x	1
Book preface				x	1
Dictionary				x	4
Bibliography				x	5
Interactive				x	2
Narrative				x	2
Book announcement				x	1
Private portrayal				x	1
Commercial portrayal				x	2
Lecture notes				x	2
Redirect				x	1

Continued on the next page

Continued from the previous page					
Discussion				x	2
Caption				x	2
Quiz				x	1

Table 5.2: The 30 fine-grained genres identified in the article class

5.4 The choice of classification models

For this work, the question of which classification model is the best for genre based classification in general, or for the actual features and genres considered in this work in particular, is a marginal issue, as has been pointed out in the opening of this chapter. There is also enough evidence that the so-called Support Vector Machines (SVM) generally outperform many other algorithms (see Section 4.4). To include that algorithm thus seems a good choice.

In addition, even though some models are considered more apt for genre classification or classification in general, any performance figures arrived at are not directly comparable to previous research results, because the corpus and/or the genre space is different. The only figures that seem relevant with respect to other research is an estimation of a base level of performance in relation to previous research. This is one more reason why SVM cannot be ignored.

It is perfectly possible for anyone with good programming skills to tailor an implementation of some arbitrary model for supervised or unsupervised learning. However, this would be a time-consuming task and since there are readily available packages that implement the models for research experimentation, it is more convenient to choose from them.

The WEKA "machine learning workbench" is chosen as the software package for the experiments. The most obvious reason is that WEKA implements several well-known algorithms, both unsupervised and supervised. WEKA is developed at the University of Waikato, New Zeeland, and is issued under a GNU public licence. It is documented by Witten & Frank (2005) in a book that also gives a general overview of data mining at large.

One of the models chosen to apply is thus the SVM. However, SVM is nothing but a generic name for several implementations that are based on one basic algorithm implemented for binary classification. Its main principles are fairly simple in theory and in some way resemble linear regression methods. If we think of the space formed by all feature-values of the data as a nonlinear multidimensional space of vectors, SVM transforms this space so that linear regression principles may be applied to derive a so-called *maximum margin hyperplane* that separates the two classes. If there are more than two classes, this binary classification process can e.g. be performed in sequences. The SVM implemented in WEKA and used for the experiment of this work is called the Sequential Minimal Optimization (SMO), originally implemented by John Platt (Witten & Frank, 2005, p. 214ff). There is a lot of parameter settings to tune, but as we want the model to behave similar to how it has been used in previous experiments, similar settings are applied here as well.

The data used in any set of experiments usually vary with respect to noise. This also holds for the data used in this work. As has been pointed out in the previous section, the `KI-04` data contains obvious misclassifications and without a tedious inspection of all the data, we cannot be sure of the quantity of this type of noise. Moreover, we do not know beforehand exactly what kind of effect noisy data have on the performance of any type of classification, given the type of task and features.

It is also relevant with respect to the experimental research question to investigate the effect of different classification models, even if we don't do an exhausting investigation. A second model is therefore chosen to be tentatively compared to the SVM model, as different feature sets and genre spaces are employed. In chosing this model, the author's previous experience of classification models has been influential and the choice fell on the relatively simple k-*NN* model, a so-called "lazy" model for *Memory Based Learning*. k-NN is a shorthand generic expression for several implementations of lazy learning, briefly characterized in Section 3.2.1. Its name is derived from the implementation of Aha et al. (1991), who showed that it is plausible to sometimes not only consider the most similar instance for class assignment, but the $k \geq 1$ most similar instances. Lazy models do not in strict terms perform any kind of learning. The instances that are to be classified are only compared to the k most similar instances. There are more advanced implementations of this model that, for instance, employ feature and instance weighting (See, for instance, the TiMBL algorithm by Daelemans et al., 2004), but the WEKA implementation is the most simple one, with the possibility of enabling distance weighting for "class voting".

A third model, an unsupervised clustering model, is also used. Clustering, briefly described in Section 3.2.2, does not need preknowledge of class adherence in the training data. The model chosen that is offered by WEKA is the *K*-means. It is chosen because it is simple and fairly similar to k-*NN* in the sense that it is based on similarity measures. Its basic idea can be described as follows.

Given a data set $\mathcal{X}$ and a decision on how many resulting clusters K are wanted, *K*-means is given N initial seeds, which typically come from $\mathcal{X}$. All remaining instances in $\mathcal{X}$ are then each assigned to the cluster formed by the closest initial seed, and a center is then computed for each of these N resulting clusters. These new pseudo-points in instance space are taken as the N seeds for iteration of the process. Iteration then continues until there are no longer any instanceces that are reassigned to another cluster. As long as $K = N$, it is the whole story; when $K \neq N$ its behaviour needs to take into account the merging or splitting of clusters, which is a bit more complicated. The goal for *K*-means can be described as minimizing the sum of squared distances between instances within clusters. This actually means that the goal of K-means is in some sense conformant with the (theoretical) harmonic quality introduced in Section 3.1. This is the reason for choosing *K*-means. However, the sum of squared distances is bound to increase with a decreasing K, and *K*-means does not actually halt its iteration when the sum of squared distance for $\mathcal{X}$ is minimized (its global measure), but only when it holds for the clusters. It is local, because it only considers similarities within clusters, and not dissimilarities between instances in different clusters.

5.5 Feature sets and their derivation

In order to answer the question of how different document features influence classification, a large set of features need to be derived by algorithmic means. A set of small programs have been written by the author for this purpose, in the multi-paradigm programming language OZ. They can be offered on request, for anyone who wants to control the results of this study or who wants to use them for other purposes.

The principles are fairly simple. The basic procedure can be briefly described as follows.

The *KI-04* corpus is delivered as a large set of HTML files $\mathbf{D}$. An initial procedure loads the contents of the first file in $\mathbf{D}$, transforms all alphabetic characters to lower case, separates the plain text T from the HTML encoded contents H, tokenizes T, and counts occurrences of features according to the descriptions of features given below. Some counts are performed on T, while other counts are performed on H. This process is repeated for the next file in $\mathbf{D}$, and so on until $\mathbf{D}$ is empty. Some tokens or characters, such as URLs in plain text or carriage return characters, need special care.

The main drawback with this process is that each token is taken as being context-free, and nothing is recorded of its neighbouring tokens or in which position of the document it is located.

In order to assess the impact of different feature sets, which is implied in the fourth research question, they are divided into the following three subsets:

1. A base subset of features that are supposed to set a baseline for assessing the impact of the other subsets, which may be justified from previous research and linguistic theories on registers and text types, may form the foundation. This subset will be denoted as $\mathbf{x}_{base}$.

2. A subset of features that represents the occurences of speech act types. These are justified by the enumeration and typology set up by Austin (1975). This subset will be denoted as $\mathbf{x}_{act}$.

3. A subset of features that are derived from the occurrences of

genre modules (corresponding to the ideas of Rehm (2002)), as identified and annotated by means of ocular inspection. This subset will be denoted as $\mathbf{x}_{struct}$.

These three subsets comprise the full feature set, such that Equation 5.1 holds:

$$\begin{aligned} &\mathbf{x} = \bigcup\{\mathbf{x}_{base}, \mathbf{x}_{act}, \mathbf{x}_{struct}\}, \\ &\mathbf{x}_i \subset \mathbf{x} \end{aligned} \tag{5.1}$$

The following sections will enumerate and motivate these features.

5.5.1 Base features

The base set of features considered in this project can be subdivided into three broad groups, where the first group contains a large portion of the features that Biber used in his work of 1988, and of the "linguistic facets" used by Santini (2007). All of these features are based on frequency counts that have been normalized and standardized in different ways.

1. *Linguistic properties of the text*, that can be computationally expensive or cheap. Expensive properties are those that require disambiguation in the form of shallow parsing or parts-of-speech categorixation, while the cheap ones are those that can be deterministically extracted by token specification.[8] The set of features of this category will be denoted as $\mathbf{x}^{l}_{base}$.

2. *Text-grammatical markers*, a category whose individual markers are at the level of punctuation, such as the number of punctuation characters, including token and document length. The set of features of this category will be denoted as $\mathbf{x}^{t}_{base}$.

3. *Markup tokens*, which are the numbers of the HTML tags in a document. The set of features of this category will be denoted as $\mathbf{x}^{m}_{base}$.

[8] A *token* is any set of consecutive characters delimited by one or a few predefined characters, usually the whitespace character.

The idea of, for instance, frequency counts of linguistic phenomena, is that for some sets of linguistic phenomena high or low values seem to co-occur frequently. Biber (1988), for instance, identified one underlying textual dimension as a "factor" that was realized as high frequency values for infinitives, prediction modals and suasive verbs. These features were then interpreted as representing a factor of "Overt Expression of Persuasion" (Biber, 1988, p. 111). The idea of using frequency counts as indicators of underlying dimensions has been applied to other document properties, such as HTML tags and punctuation characters, in previous research on genres. Previous research on genre and register variation is important to consider when choosing features for classification, in order to avoid overestimated values for certain underlying textual dimensions.

A reasonable and convenient way to compensate for any bias towards certain underlying dimensions is to apply clustering techniques or simple covariance measures on features before using the features for document clustering. On the other hand, if the data samples are restricted and the real task for which the application is expected to be used is far more extensive, then such methods may result in overfitting problems. That is, the tuning of a classification model becomes too much a product of the size and nature of the sample data.

The linguistic features consist mostly of lexical categories based on approximations of their possible linguistic functions. In some cases, this may have the consequence that a certain word falls in two categories, meaning that it is being counted as an occurrence of two different features. This happens because the features are not mutually exclusive. The selection is guided by, on the one hand, the enumeration by Biber (1988, p. 223ff) in his early register studies, and, on the other hand, the features presented in the recent thesis by Santini (2007).

Both Biber and Santini base their works on the existence of more or less complex linguistic pre-processing. Biber's work is carried out on corpora with words tagged with linguistic categories. In the case of Santini (2007, Chap. 5), syntactic and "functional" patterns, whose identification relies on the NLP tool CONNEXOR, are used for the identification of a set of abstract categories that she terms "facets".

Such tools are avoided in this work for the reason of computational simplicity, which will probably make the linguistic feature set less accurate.

The feature derivation process relies on the naÃ¯ve assumption that a simple occurrence of a certain lexical item indicates the occurrence of a linguistic category. This is an obvious drawback, since tokens, i.e. a sequence of characters delimited by whitespace, may be constituted by different words or signs of any kind. For instance, the token ’me’ can be a personal pronoun but can also be used in conjunction with ’windows’, in which case it is most probably a reference to the operating system Windows ME — i.e. an abbreviation. The token ’promise’ can be both a public verb and a noun. Without NLP tools that detect linguistic categories and relations, such ambiguities remain unresolved. Either such ambiguous tokens are excluded or the ambiguity is ignored. If it is ignored, it is for the sole purpose of making the extraction of linguistic features as computationally inexpensive as possible, at the cost of a decreased accuracy in feature derivation.

This level of decreased accuracy has not been investigated. Its potential negative effects have been managed by discarding some of the lexical items enumerated by Biber and Santini. This goes for the personal pronoun, first person singular, in the nominative case (’I’). A choice that increases precision, but decreases the recall of this feature detection. This simplified approach to feature derivation also affects those categories that include phrases. Santini (2007, App. B) lists ’by the way’ as an indicator of “discoursal connective”, but since this phrase requires the parsing of sequences of tokens, it is discarded from the count.

Another unfortunate consequence is, of course, that despite the removal of obvious ambiguities, some features may reflect two or more underlying linguistic functions. If ’by the way’ is included, for instance, it can be a discoursal connective, but also part of a prepositional phrase signifying the manner in which something is done, such as in ’by the way of an example’. However, the aim of including linguistic features is only to establish a reasonable baseline, and hopefully these shortcomings will not affect the investigation too much.

Here follows the enumeration of the features. Each feature is given a numeric reference $x_i, i = 1, ..., 55$, and a more mnemonic label used in the experiments.

Pronouns

Pronouns are in themselves undetermined references and, except for indefinite pronouns, need to be resolved by the reader through the cotext or the situational context (Biber et al., 1999, p. 70). The choice of pronominalization, especially with respect to personal pronouns, may indicate the stance of the author taken towards what is being said.

Feature x_1 First person pronouns singular (pronouns_1st_sing)
The frequency of the first person pronoun in singular form may indicate a degree of personal involvement that serves to distinguish between e.g. private portrayals and academic prose. However, 'I' is excluded because it is ambiguous and may often be used in roman numericals of enumerations of different kinds.

The feature includes the following words:
{me, mine, my, myself}

Feature x_2 First person pronouns plural (pronouns_1st_pl)
The plural form of the first person pronoun is in some sense a parallel to the singular form, but there is an interesting observation that has often been made. Dura et al. (2006) report that in biomedical texts, the first person pronouns in the plural form make up 50% of the occurrences of pronouns, which should be compared to "general prose", in which it is 5%. When research teams within certain scientific disciplines report research activities, they use them extensively, while it may be different for other scientific genres.

The feature includes the following words:
{we, us, our, ours, ourselves}

Feature x_3 Second person pronouns (pronouns_2nd)
This feature may indicate that the author(s) urge(s) the reader to pay

special attention and react to something expressed in the text. Thus, it often marks a direct invitation to the reader common for instructive texts. Biber also attributes its use to narrative texts.

The feature includes the following words:
{you, your, yourself, yourselves}

Feature x_4 Indefinite pronouns
The indefinite pronouns realize undetermined references. Their function is not particularly clear, but the feature is included in Biber's first factor, "involved versus informational production".

The feature includes the following words:
{anybody, anyone, anything, everybody, everyone, everything, nobody, none, nothing, nowhere, somebody, someone, something}

Adverbs

Adverbs may function as either modifiers or adverbials (Biber et al., 1999, p. 538ff). In many cases they indicate a particular stance taken by the author, or serve as a linking facility between clauses. Often, they are not necessary for a linguistic expression to mediate a statement, but rather indicate a particular compositional strategy that in its turn can be indicative of some particular text type, in the sense of text types that Werlich (1976) apply.

It is true for adverbs that the same adverb may have different functions in different cotexts. Thus, as is the case for most of the features, categories of adverbs can only be identified with a certain unknown probability of accuracy.

Feature x_5 Amplifiers (ampl)
Amplifiers are the opposite of downtoners, but their covariance with downtoners is non-significant, according to Biber's investigation. They occur in more informal texts.

The feature includes the following words:
{absolutely, altogether, completely, enormously, entirely, extremely, fully, greatly, highly, intensely, perfectly, stringly, thoroughly, totally, utterly, very, leading, significant}

Feature x_6 Downtoners (dwntone)
These adverbs indicate, for instance, that the author wants to downtone the certainty or degree of a statement. It is expected to demonstrate a degree of uncertainty sometimes considered scientifically correct.

The feature includes the following words:
{almost, barely, hardly, merely, mildly, nearly, only, partially, partly, practically, scarcely, slightly, somewhat}

Feature x_7 Spatial adverbials (spat_adv)
The descriptive text type, mentioned by Werlich (1976), would typically exhibit many spatial adverbials, as it is concerned with spatial phenomena. Biber stresses that both spatial and temporal adverbials mark a concrete theme for the texts, as opposed to an abstract one. Travel reports may typically use many spatial adverbials. In that way, this feature may also indicate narratives.

The feature includes the following words:
{aboard, above, abroad, across, 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}

Feature x_8 Temporal adverbials (temp_adv)
What has been said of spatial adverbials may also be appropriate for temporal adverbials, even though they are used in the context of referrals to phenomena and events in time, and thus are typical for narratives.

The feature includes the following words:
{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}

Connectives

Connectives adhere to a group of expressions that are mostly realized by means of adverbs or adverbials. They are often used to link parts of the text with each other and create a textual coherence. Werlich (1976, p. 202f) connects the different categories to the realization of different text types.

Feature x_9 Enumerative connectives (enc)
The feature includes the following words:
{finally, firstly, secondly, lastly, thirdly}

Feature x_{10} Equative connectives (eqc)
The feature includes the following words:
{correspondingly, equally, likewise, similarly}

Feature x_{11} Reinforcing connectives (reic)
The feature includes the following words:
{besides, furthermore, moreover}

Feature x_{12} Appositive connectives (apc)

The feature includes the following words:
{namely, specifically, viz}

Feature x_{13} Resultative connectives (resc)

The feature includes the following words:
{accordingly, consequently, hence, somehow, therefore, thereby}

Feature x_{14} Inferential connectives (inc)

The feature includes the following words:
{otherwise, else}

Feature x_{15} Summative connectives (suc)

The feature includes the following words:
{thus, overall}

Feature x_{16} Reformulatory connectives (refc)

The feature includes the following word:
{alternatively}

Feature x_{17} Antithetic connectives (anc)

The feature includes the following words:
{contrariwise, conversely, instead, oppositely}

Feature x_{18} Consessive connectives (conc)

The feature includes the following words:
{admittedly, anyhow, anyway, anyways, besides, however, spite, nevertheless, nonetheless, notwithstanding, yet, though}

Feature x_{19} Discoursal connectives (disc)

The feature includes the following word:
{incidentally}

Feature x_{20} Temporal connectives (tec)

The feature includes the following words:
{eventually, meantime, meanwhile, originally, subsequently}

Verb categories

Verbs may be categorized in many different ways. Santini (2007) used a fairly fine-grained semantic scheme borrowed from the *Longman Grammar of Spoken and Written English* (1999) and Biber (1988) used a broader scheme based on Quirk et al. (1985). What is wanted in the context of genre investigations are categories of verbs that may be mapped to different kinds of common tasks, situations, and communicative expectations in which they frequently or rarely occur. The choices made in this work are not based on an independent and encompassing survey of linguistic literature, as that would be too far-reaching. The verb categories are borrowed from the investigations of Biber (1988) (public, private and suasive verbs).

Feature x_{21} Public verbs (pub_vb)

Public verbs is a class of verbs that most often introduce "indirect statements" (Quirk et al., 1985, p. 1180). Semantically they refer to actions that are supposed to be observable. They are used frequently in narratives (Biber, 1988, p. 108f).

The feature includes the following words:
{acknowledge, admit, agree, assert, claim, complain, declare, deny, explain, hint, insist, mention, proclaim, promise, protest, remark, reply, report, say, suggest, swear, write}

Feature x_{22} Private verbs (priv_vb)
Private verbs is a class of verbs that refer to “intellectual states such as belief and intellectual acts such as discovery” (Quirk et al., 1985, p. 1181). They are “’private’ in the sense that they are not observable”. In Biber’s work, private verbs co-occurred with the frequent use of the first person pronouns, marking texts with an intention of “overt expression” (Biber, 1988, p. 105).

The feature includes the following words:
{anticipate, assume, believe, conclude, decide, demonstrate, determine, discover, doubt, estimate, fear, feel, find, forget, guess, hear, hope, imagine, imply, indicate, infer, know, learn, mean, notice, prove, realize, recognize, remember, reveal, see, show, suppose, think, understand}

Feature x_{23} Suasive verbs (suasives)
Suasive verbs should, according to Quirk et al. (1985, p. 1182), be juxtaposed with “factual verbs”, which is a general category that includes private and public verbs. Suasives generally perform directives. As the label for this verb class indicates, suasive verbs often mark an intention of the author to persuade the reader.

The feature includes the following words:
{agree, arrange, ask, beg, command, decide, demand, grant, insist, instruct, ordain, pledge, pronounce, propose, recommend, request, stipulate, suggest, urge}

These are all word occurrences used by Biber. Among those omitted are, apart from those that require expensive natural language processing, lexical items that are ambiguous, such as ’I’ and ’that’, and lexical items that are compounds of two words, such as the conjunct ’for example’.

Text-grammatical features

The following features belong to the text-grammatical category. They typically serve the purpose of either delimiting, separating, or distinguishing a low-level text constituent that may correspond to elements of a document structure — a text unit. They are termed text-grammatical in the sense of Nunberg (1999), even though the document length and the average token length cannot rightfully be characterised as text-grammatical measures. One of the advantages of this category is that it probably constitutes the most computationally inexpensive set of features that is to be extracted. One of its disadvantages is that these features are not functionally homogeneous and that some of them may be extensively used for highly different purposes.

Feature x_{24} Document length (doc_lengthT)
Document length, measured in number of tokens or characters, is used as a normalizer to balance frequency counts. However, it is clear that some broad types of texts are frequently long or brief, which motivates the investigation of document length as a feature in itself. As a feature, document length is always counted on the basis of tokens, not of characters.

Feature x_{25} Average token length (token_length)
High frequencies of this feature indicate lexical complexity. It should be noted that this feature is not a pure linguistic feature, as it is contaminated by the fact that tokens are not necessary occurrences of word forms. They may be, e.g., URLs, numbers, or any arbitrary sequence of characters.

Feature x_{26} Colons (colons)
Colons often follow upon an introductive expression indicating that what follows is a consequence, introducing a following description, a quote, an enumeration, or something else that needs to be thoroughly introduced. In addition, the colon is used for many other purposes in

e.g. mathematics and bibliographic descriptions.

Feature x_{27} Commas (commas)
Commas are usually separators that break down complex expressions into smaller units. Thus, their frequency can be considered typical for complex sentences.

Feature x_{28} Exclamation marks (excl_marks)
An exclamation mark typically ends an imperative expression of some kind. Formal texts are typically expected to lack exclamation marks, unless it includes much programming code, is a mathematical text, or some other text that makes use of punctuation characters in an artificial language.

Feature x_{29} Question marks (qu_marks)
Question marks are designed for marking questions, and may thus indicate expressions that are more informal towards the targeted reader. So-called FAQs (Frequently Asked Questions) would likely contain many question marks.

Feature x_{30} Quotation marks (quot_marks)
Double quotes may delimit quotations or mark words that are to be understood in a diverging fashion. They are also extensively used in some bibliographic styles. Quotation marks have different graphical appearances in different contexts and in different languages. Single quotes are frequently used and are included here. However, the restricted character set on the web makes it fairly likely that even the French style for quotation marks has been avoided.

Feature x_{31} Full stops (stops)
Full stops are separators of smaller text constituents and perform a similar (disambiguous) function as commas do. However, they are also used to indicate abbreviations and occur in several kinds of enumerations.

Other punctuation characters possible are discarded because of their heterogeneous functions. Semi-colons are, for instance, discarded for technical reasons, since they are extensively used as escape characters in external entity references.

Markup token features

The following markup encodings are used as features:

Feature x_{32} Anchors (anchors)
Anchors are the bearers of hyperlinks and rather special for web documents. They are in no way unambiguous with respect to their communicative function. Anchors may be included only by convention to support navigation. In this case, they provide a technical function which has almost no bearing on the contents. In other cases, anchors perform a similar function as when bibliographic references are included in printed texts to support argumentation, or to pay respect to someone else's opinions or works.

Feature x_{33} Form elements (form)
Form tags are used to embed interactive forms, such as the frame for user name input. In addition, authors may choose to embed a form for input in which the user may enter a search query for a database engine.

Feature x_{34} The ratio of paragraphs and headings (ph_relation)
`p`-tags are intended to be used for dividing a piece of text into paragraphs, and `h1`, `h2`, and `h3` as headings for sections of the text. A conventional academic text generally has a certain balance between the number of paragraphs and the number of headings. Headings should be fewer than the number of paragraphs, which is probably not be expected for a text which is much shorter and intended to attract the user's attention by typographical means. However, the purposes of the tags are often violated. For instance, the `font`-tags may be used instead of the heading tags.

Feature x_{35} Metadata elements (meta)

The empty meta-element of the HTML standard is often indicative of a concern for the document being discovered by different kinds of search engines. It is likely to mark some kind of extended care that the text is being appropriately described and retreivable, although it may also only convey which character encoding scheme that has been used, or other technical information automatically added by a dedicated HTML editor.

Feature x_{36} Image elements (img)

The `img` tag is used to embed images of different kinds in HTML documents. It is plausible to assume that the inclusion of images in academic writings is generally for illustrative purposes and motivated by the content, and that, for certain domains, the frequency of embedded images is probably lower than for a web shop announcement.

Feature x_{37} List items (li)

List item tags mark items in either an ordered (and numbered) list, or an unordered itemized list. Thus, it may functionally resemble the linguistic category of enumerative connectives in establishing textual coherence of a particular kind (Cf. Biber et al., 1999, p. 875ff).

Feature x_{38} Pre-formatted areas (pre)

The `pre`-tag is often used for the insertion of preformatted text directly into a web page, with no other markup than whitespace characters, linebreak characters, and interpunction. This is typical for e.g. sequences of programming code. At other times it is used in order to mimic type-writer style, or for tabular data, when the author considers the effort of encoding tabular data as an HTML table too large.

Feature x_{39} HTML tables (table)

HTML tables are defined by the opening and the closing `table`-tags that surround a set of other tags used for the specification of table rows, cells, and headers. Tables are of course used for encoding table data, but frequently also for formatting purposes, when the author wants to adjust page margins or define a document as a document of n number of columns. Thus tables mark two very different functions of the enclosed text.

5.5.2 Speech act features

As has been pointed out in Chapter 4, speech act categories have been the focus for at least one other case of genre studies, although exclusively for e-mail categorization (Goldstein & Evans Sabin, 2006). The rationale behind using speech acts as genre indicators can be exemplified by the exercitive "I advise", which would be awkward in a scholarly paper, but not unexpected in a discussion on an on-line forum. Austin lists five categories of verbs whose occurrence (might) signify certain kinds of performative character of an utterance in which the verb acts as a finite verb. However, according to the speech act theory, verb forms in present tense are the only occurrences that are counted. An expression such as 'I define X as Y' obviously has an illocutionary force, while an expression such as 'he defined X as Y' is of a reporting character and does not necessarily reveal anything about the personal, situational, or interpersonal aspect of a speech act. Past tense verb forms are included in the base set, but not as tokens of the speech act features.

Two important facts must be observed with respect to these features. First, many of the verbs depend on their grammatical cotexts in order to properly be assigned to a category and are only in themselves possible indicators of the property assumed for that category. In order to assign each occurrence of a verb to its correct category, more complex natural language processing would be required. For instance, the exercitive 'command' cannot be distinguished from the noun 'command', or from when it is used in a prepositional phrase such as '... in the way we command'. This is obviously a great disadvantage and should be remedied in further experiments. Second, regarding Austin's categories, they are, as he has also emphasized, sometimes not clearly exclusive. Some verbs may be assigned to several categories. This means that each of these features includes several verbs that are to be included in another feature count as well, in the base subset as well as in this subset. For instance, the token 'promise' is thereby counted as both a public verb (of the base subset) and a commissive verb.

A final remark can be made on Austin's catalog. It dates back to the 1950s and reflects a vocabulary in which some verbs seem fairly unusual today. The choice, for this study, is to remain conservative. If some verbs do not occur at all in the texts, because they are archaic, it will not make any difference. the conservative approach is disadvantageous only if important verbs of a certain character are missing, due to their recent inclusion in everyday language. Further research may refine this catalog in several ways, if the features show any sign of being valuable.

Feature x_{40} Behabitives (behab)
Behabitives express speech acts that relate to how other people behave or react (Austin, 1975, p. 160).

The feature includes the following words:

{apologize, thank, deplore, commiserate, compliment, condole, congratulate, felicitate, sympathize, resent, pay, tribute, critisize, applaud, overlook, commend, deprecate, blame, approve, favour, welcome, bless, curse, toast, wish, dare, defy, challenge}

Feature x_{41} Commissives (commiss)
Commissives signify commissions of the author to act in certain ways in the future (Austin, 1975, p. 157ff).

The feature includes the following words:

{covenant, contract, undertake, intend, purpose, contemplate, envisage, engage, guarantee, bet, vow, consent, adopt, champion, embrace, espouse, oppose, favour}

Feature x_{42} Exercitives (exerc)
Exercitives are similar to commissives, but they commit the *reader* to act in certain ways (Austin, 1975, p. 155ff).

The feature includes the following words:

{sentence, fine, levy, vote, nominate, choose, give, bequeath, pardon, resign, warn, plead, entreat, press, announce, quash, countermand, annul, repeal, enact, reprieve, veto, dedicate}

Feature x_{43} Expositives (expos)
Expositives are often used in argumentations for in order to clarify "reasons, arguments, and communications" (Austin, 1975, p. 163).

The feature includes the following words:

{affirm, state, describe, identify, remark, interpose, inform, apprise, tell, answer, rejoin, testify, conjecture, accept, concede, withdraw, demur, adhere, repudiate, correct, revise, postulate, deduce, argue, neglect, emphasize, interpret, distinguish, analyse, analyze, define, illustrate, formulate, refer, call, regard}

Feature x_{44} Verdictives (verd)
Verdictives exercise some kind of judgment that relates to truth or falseness, or good or bad (Austin, 1975, pp. 153,163).

The feature includes the following words:

{acquit, convict, hold, interpret, rule, calculate, reckon, locate, place, date, measure, grade, rank, rate, assess, value, describe, characterize, diagnose, analyse, analyze}

5.5.3 Document structure features

The logical document structure, as defined in Chapter 2, Section 2.3.3, has been treated by both Rehm (2002) and Mehler (2007) as an important (but "uncertain") characteristic of web genres. Rehm does not explicitly elaborate its applicability to genre classification and has a somewhat broader aim, weheras Mehler is mainly concerned with the modelling of larger coherent entities than is the case here. Mehler et al. (2007) shows that the logical document structure is in itself an effective source of features that perform well for genre classification. However, the experiments of Mehler et al. are applied to a fairly homogenous set of news material in which, as it seems, the use of XML markup is fairly valid and predictable.

The logical document structure in this work is considered to be conveyed by the use of the rather unpredictable HTML markup encountered on the web, resulting in structures that are highly ambiguous with respect to what kind of document structure constituents they embody. In order for the markup to be used as a source for feature derivation, it has to be disambiguated in some way, which in itself constitutes a classification problem.

The decomposition principle presented in Section 2.3 technically defines a document as a structure of text nodes. From the perspective of the logical document structure, a document is thought of as a structure of genre modules. The classification task $\mathcal{T}$ here can thus be defined as in Equation 5.2, where $\mathcal{N}$ is the resulting set of nodes of a document after its decomposition and $\mathcal{G}m$ is the set of possible genre modules defined.

$$\mathcal{T} : \mathcal{N} \rightarrow \mathcal{G}m \tag{5.2}$$

When such a disambiguation is accomplished, a document may be represented as a structure of genre module occurrences. The complexity of this classification task is considerable and thus computationally expensive. However, in some preliminary investigations for this work, this text node classification has been tested on a restricted sample of reviews and ordinary articles from two different electronic journals. The complete data set consisted of over 5 000 text nodes, and an accuracy around 70% was attained by a simple probabilistic classifier. Whether such a fairly expensive preprocessing issue is worth the while to investigate further, can hopefully be estimated by the results of this work.

In this work a pragmatic stance is taken, and in order to test this kind of features a manual annotation has been undertaken, in which for each document the number of occurrences of a set of genre modules has been counted and recorded in the header of each document in the *articles*-class. The genre modules counted will be enumerated below.

The features of the subsets $\mathbf{x}_{base}$ and $\mathbf{x}_{act}$ are taken as counts of their occurrences. These features may take on any value from the set of real numbers. Even though these values are transformed and restricted in a normalization process, the difference between their raw frequencies is still significant. For many document structure features, this is not the case. For instance, the difference between zero occurrences and one occurrence of a list of references is more significant than the difference between two and four occurrences.

Feature x_{45} Responsibility statements (resp_state)
When documents are given titles, this is often done in conjunction with a declaration of who is responsible for the contents of the document. A responsibility statement is then a declaration of the author or corporate body responsible for the contents of a document, possibly together with data associated with the author or corporation, such as affilitation and e-mail adress, as well as the title. Functionally, such statements are imperatives, serving the purpose of making later claims of intellectual property possible. A responsibility statement often obeys the conventional structuring of certain editorial guidelines and is included more or less by routine.

In this functional view, a copyright statement could be counted as a responsibility statement. However, a copyright statement does not necessary have anything to do with the intellectual contents, but with the document actually being made public.

In most cases, a responsibility statement occurs in close connection to the title statement at the beginning of, and in a prominent position of, a document. If it does not and it is clear that the name of the author refers to the person who has made an intellectual effort of writing or artful production, it is still counted as a responsibility statement. Certain documents may be compilations of several contributions in one single file, and thus each contribution may have its own responsibility statement.

Feature x_{46} List of references (reflists)
On the surface, a list of references is an enumeration of documents by way of more or less descriptive identifications. Academic styles require that these descriptions are standardized, given titles, names of corporate bodies or individuals responsible for the content, year of

publication, publisher etc. In other cases, they may only consist of an identifier, such as a URL.

Functionally, lists of references may be interpreted differently. In academic contexts, they usually refer back to inline citations of the main text that may support an argumentation or just serve to establish a relationship between the contents and other documents or authors. On a personal homepage, the listing may be intended to demonstrate what the object of the personal homepage has written or it may declare his or hers interests. In the *KI-04* collection, there is one category labeled "link lists", whose main contents are lists of references that, unlike lists of references in typical academic contexts, are not linked to citations in a main text. These "link lists" functionally resemble bibliographies in that they enumerate documents and/or give guidance on what to read on a particular topic.

Another difference between various lists of references is what kind of documents they refer to. For instance, the document on the *Levenshtein Distance* (article_4225846718) includes a section of references that refer to program code, other documents refer to particular software by way of pointing to its online documentation, still others refer to the points of entrance for web sites of organisations. These last examples are clearly functionally different from a list of references in a typical argumentative or expository research article.

In addition, the documents referred to may be closely related to the document within which the list of references occurs. For instance, the CV of a personal home page may list the production of an author and be considered navigational support.

A list of references is, as pointed out above, sometimes an enumeration of citations made in a text. This does not mean that the citations are part of the list of references. The citations establish document relations *by means of* the list of references. Citations, however, may refer directly to other documents within running text or in a footnote or endnote, with more or less elaborated references. Thus, citations may be functionally similar to a list of references, but since they are not collocated and graphically set apart from the surrounding text, they are not counted as lists of references. The reason for this, is that the problem of algorithmically identifying occurrences of citations is particularly difficult, especially since their formal realizations vary and they often do not constitute a distinguishable element by means of surrounding markup.

The genre module of lists of references is defined on a level of granularity that counts enumerations of references to documents as lists of references provided that they 1) are graphically set apart from the surrounding text, 2) refer to documents that cannot be considered parts of or highly related to the document in which the references occur (in this case they are treated as navigational items), 3) are given a descriptive element in addition to being an identifier (i.e. not only a list of linked URLs).

Sometimes a list of references, and which is also interpreted as one, may be structured by the help of headings. In this case, the headings are considered as dividers and each heading opens a new list of references.

Feature x_{47} Quotations (quotes)

A quotation is a longer quote from another source. It is typically set off from the rest of the text and most often included in a contents section, even though it is fairly common that it follows directly upon a title and responsibility statement or a heading. Shorter quotations that are not expressed as independent blocks, such as inline quotes marked by citation signs, are ignored. Their functions are not uniform. They may be illustrative or emphatic (supporting an argument).

Feature x_{48} Abstracts (abstracts)

An abstract describes and summarizes (sometimes very briefly) the contents of the document or, if the contents are distributed over several files, the contents of several documents. It has a descriptive function that is more encompassing and elaborated than a title or a heading. It is typically marked by occupying a position directly before the main contents and by being directly preceded by a title and a responsibility statement. In academic genres, such as various kinds of journal articles, abstracts often summarize the research objectives, the methodology, and the findings. In other cases, one may encounter introductory text constituents that do not give such accounts, but instead describe the target audience or the disposition of the text. If this is done in an elaborated way, it is counted as an abstract.

On the web, it is not uncommon for a document to only consists of an abstract, a title, and a responsibility statement, where the document functions as a starting point for the reading of a larger physically fragmented but intellectually coherent text.

Any coherent text constituent that describes the following text, its disposition, its intended use, or its history is considered an abstract.

Feature x_{49} Notes (notes)

A note is a kind of marginal comment or elaboration, linked to some point in the contents but not necessarily by technical means (i.e. by way of HTML hyperlinks). It has a function of marginal support and elaboration to the contents section. Footnotes, endnotes and marginal comments are treated as equals. However, sometimes a piece of text labeled "note" may appear, which is netiher a comment nor an elaboration on the text, but intended to make the reader to observe something particular. Such constituents are not counted as notes.

Feature x_{50} Figures (figures)

A figure is any material that is most often intended to illustrate the contents, furnished with a caption that generally gives the figure a unique identification within the text and/or a piece of text that functions as a description of the figure. In the DocBook DTD, a figure element may wrap many different kinds of illustrative material, such as program listings and "informal equations", and is required to have a caption.

The genre module is more loosely understood, but also narrowed down to include graphics only. That is, equations and program listings are not counted, as they rarely have a caption and are therefore difficult to distinguish from inline equations, or inline mathematical or formal expressions. If captions or another descriptive label for graphics is absent, the item is not counted as a figure unless the figure is explicitly referred to in the preceding or succeeding text. That is, the graphical material needs to be a clear part of the text. A figure can consist of ASCII characters embedded in an HTML PRE element, such as in Figure 5.2, and is then counted as a genre module.

Sometimes, several embedded graphic files may have one single caption, describing a sequence of illustrative graphics. In this case, they are counted as one occurrence of a figure. Figures may be represented by a sole caption and a hyperlink to an external file with graphics. Such occurrences are not counted. A particular problem occurs when tabular data are mistakenly labeled as a figure and figures contain text or equations represented as graphics. In the first case, it is not a figure, and in the second case, it is a figure if it is explicitly referred to as a figure.

```
1.0 +                   +-------------------
    |                  /
    |                 /
0.5 +                /
    |               /
    |              /
0.0 +-------------+-----+-------------------
                  |     |
                 5.0   7.0
```

Figure 5.2: ASCII figure

Feature x_{51} Tabular data (tables)
The function of tabular data in connection with the surrounding text is very similar to a figure, but tabular data are perfectly ordered into rows and columns that structures textual or numerical data. HTML tables are frequently used on the web for text layout purposes only, but are then consequently not counted as tabular data. Tabular data may sometimes be layed out in an HTML PRE element by using series of whitespace characters to simulate cell borders. This is similar to how figures may be drawn, or how a simple enumeration may be realized and makes it somewhat difficult to determine when there is a table, a figure, or some other kind of structured data. However, the tabular data

feature is required to be structured in a way where rows and columns contain data that are typologically consistent.

Feature x_{52} Navigational block (navigation)
A navigational block forms a piece of mechanical support for moving *within* a document or a web site, primarily by means of hyperlinks. Tables of contents without mechanical support (i.e. without HTML hyperlinks) are treated in the same way. On the web, the use of hyperlinks is a prominent feature and many texts employ an extensive use of hyperlinks in different ways. However, only sets of hyperlinks or content listings that are clearly set apart from the rest of the text are counted as navigational blocks. An HTML form that provides the possibility for searching is considered a navigational block.

To distinguish between a navigational block and a list of references can sometimes be very difficult. If, on a personal homepage, the author links to his own works (papers, conference presentations etc.), or, if another kind of document contains a link to a printer friendly version of the same document; are these to be counted as navigational support or as external references?

The principles for resolving such ambiguities are as follows. To be counted as a navigational block, 1) the item needs to be set apart from the surrounding contents, 2) at least one item needs to refer to a particular place in the document (file) within which the item is located, or a document that is highly related to the contents[9] of the document within which the item is located, and 3) if the item(s) provides navigational support but match the criteria for any other genre module, the other genre module takes precedence.

Thus, a link to a printer friendly version of a document is a kind of navigational support, but links to the author's writings is a list of references, as they are usually elaborated bibliographic references.

[9] This requirement rules out e.g. links to the home page of the author, if it is not a link from a CV to the point of entrance of the author's web site.

Feature x_{53} Copyright statements (copyright)
A copyright statement is a claim made by an individual or a corporate body for the intellectual property of the contents , by virtue of his, her or its name. It may seem that a copyright statement is indistinguishable from a responsibility statement. However, a copyright statement does not have to be based on the fact that making the document public implies an intellectual effort of a high degree by a named person or organisation. Moreover, copyright statements are not graphically emphasized, as responsibility statements usually are. They are more often small remarks towards the end of the document and positionally separated from the title declaration.

A copyright statement is often collocated with a time stamp of some kind and a copyright sign. Thus, copyright statements usually give declarations of the circumstances around the publication. An item such as a simple declaration of the "latest update" without a reference to the body or an individual responsible for the publication is not counted as a copyright statement. However, a copyright sign together with a time stamp is considered a copyright statement.

Feature x_{54} Acknowledgements (acknowledge)
An acknowledgement is an explicit act of thankfulness to someone or something. The statement must be distinctly set apart from the rest of the text in order to count as an acknowledgement.

Feature x_{55} Definitions (definitions)
Definition lists are e.g. glossaries. They consist of lists of term-definition pairs, which should be understodd in a wider sense of the words term and definition.

Considering this feature set, it may become clear that it is highly tailored to what genre modules that can be expected in scholarly material, and not at all applicable to certain other kinds of documents, such as those found in the *KI-04* classes *discussion pages* and *download pages*. This is a conscious choice. The *articles* class contains the kind of documents that has been declared as the most interesting for this work. Extending this feature set to include genre modules of importance to the other *KI-04* classes would make the introduction of modules such as *C.V.*, *discussion postings*, *questions*, and *answers* almost necessary. There are two drawbacks to such a thing, besides the amount of tedious work needed to annotate all classes. First, the number of genre modules would extend a reasonable level for which text node classification would be plausible. Second, as some modules would be specific for some classes, they would also be enough in themselves to determine the class. This goes for *discussion postings* and probably also *C.V.*. It is then quite possible that the entire task of genre classification could be accomplished by simply training a classifier to recognize the occurrence of certain genre-specific modules. That is a task worth pursuing in further research.

Thus, we assume that in real world applications, there are robust ways to filter out scholarly material with high precision from other stuff, without the use of document structure features, and to apply document structure features to refine the classification of identified scholarly material. This is the assumption applied here.

5.5.4 Estimating the discriminative power of features

The discriminative power of features may be measured either in order to assign weights to each feature or to eliminate features that are useless or considered noisy. In this work, it is used for the purpose of elimination. If there is some pre-knowledge of a set of documents, i.e. if it is known that one subset belongs to one genre and another to some other genre, there are valuable prerequisites for estimating the discriminative power.

One possibility is to use the so-called *Z-score* for a subset that is treated as a sample from the complete data set. It can be assumed that if the mean value of a feature over a subset (a class) increases above or decreases below a certain threshold in difference from the mean value of the feature over the complete set of documents, this feature demonstrates a significant deviance from the norm, and could be judged discriminative.

However, this measure involves the mean value and is thus only considered robust for feature-values that can be approximated as Gaussian distributed. This is most often not the case for the kinds of features employed in this work. Karlgren (2000, Chap. 10) was particularly careful with such assumptions, and a quick examination of the data set of this work seems to confirm this. Such an assumption is possibly true only for average token length, whose mean and standard deviations within the eight classes are shown in Table 5.3.

If Gaussian distributions are not at hand, and as long as there are training data for which we have class labels, there are more possibilities of estimating the discriminative power of features. Three measures are commonly used for this purpose: Information Gain, Gain Ratio and χ^2. Sebastiani (2005, p. 116) refers to these as instruments for dimensionality reduction or feature selection, while they are used for feature weighting in TiMBL (Daelemans et al., 2004, pp. 20-22).

The idea behind these measures is that given a class and some kind of discretization of feature-values, class and feature-value probabilities can be used to measure the impact of a certain feature. These measures are all going to be used for the analysis of some of the features in this work.

5.5.5 Standardization and normalization of document features

The raw counts derived from the initial processing of documents form the foundations for all of the features. However, let us think of the derived data in terms of a two-dimensional matrix and consider a simplified example where there are four documents and only two features, the frequency of lists of references and the frequency of commas ($\mathbf{x} = \langle x_1, x_2 \rangle$), and that we have the following figures:

	x_1	x_2
d_1	1	0
d_2	0	10
d_3	1	40
d_4	1	40

	***KI-04* Category**							
	articles	discussions	priv. portr.	help	linklists	downloads	non-priv. portr.	shop
$\overline{x}$	4.83	4.22	5.01	4.65	5.34	4.76	5.00	4.54
σ	0.63	0.65	0.65	0.48	0.67	0.49	0.57	0.60

Table 5.3: Mean and σ for average token length in the 8 *KI-04* classes.

It is evident that the values of the two features in this example have completely different *variances*. One may also assume that the *scales* and *types* are different. A document with 60 lists of references is almost unthinkable, and (as has been mentioned above) the difference between 0 or 1 occurrence is probably more significant than that between 2 and 3 occurrences. In addition, the frequency of commas needs to be related to the length of the document. Other feature frequencies may also be dependent on their collection frequencies — the more common a feature is in the collection as a whole, the more frequent it is expected to be in one document.

With this matrix in mind, there is both a need for harmonizing the values on a horizontal axis (between features) and on a vertical axis (between instances). The former harmonization is often referred to as standardization and the latter as normalization.

The following principles for standardization and normalization have been applied on all linguistic and text-grammatical features (except of course for average token length). First, both document length and collection frequency are taken account of. The mean document length in the collection is multiplied by the raw value for feature j in document i, and divided by the document length for document i. This value is logarithmically transformed in order to diminish the effect of outliers, before 1 is added in order to avoid negative values. This part of the equation reflects a value that is normalized according to document length, and then multiplied with a logarithmic transformation of the collection frequency for feature j. We denote this measure given in Equation 5.3 as x_{ij} for feature j and document i

$$x_{ij} = \left(\left(\log_2 \frac{\overline{DocL} \times x_{ij}}{DocL_i} \right) + 1 \right) \log_2 cf \qquad (5.3)$$

where cf is the count of feature x_j in the collection as a whole, and $\overline{DocL}$ is the mean document length in the collection. $DocL$, as well as $\overline{DocL_i}$, is measured in the number of tokens for all lexical and markup based features, except for the character features ($x_{31} - x_{36}$). This may be seen as a kind of feature weighting and is an adoption of term weighting as it is applied in many information retrieval applications (cf. e.g. Manning & Schütze, 1999, pp. 541ff).

This measure is then standardized to fit between 0 and 1, according to Equation 5.4 that rescales the source measure with respect to the range of values for feature j in the collection.

$$\hat{x}'_{ij} = \frac{x_{ij} - x^{cf}_{j_{min}}}{x^{cf}_{j_{max}} - x^{cf}_{j_{min}}} \tag{5.4}$$

For some markup features and document structure features Equation 5.3 makes less sense.

1. HTML meta elements are used in the production of a document either because the encoder has a concern for document description and bibliography, or because the software used automatically adds some meta elements. Just as there is no reason to expect more title declarations for a larger document, there is no reason to expect more meta elements for a larger document. However, they cannot appropriately be interpreted on an ordinal scale, as the significance of the number of meta elements seems unpenetrable.

2. HTML form elements indicate some kind of interactive purpose, such as the inclusion of a possibility to search a web site or to log in with a user name. It is fairly common to include just one form element that directs queries to Google or any other search engine, but less common to include more than one form element. Therefore, the frequency of form elements seems to be best regarded on a three-level scale: absence, one form element, or more than one form element.

3. The ratio of paragraph tags and heading tags is probably not independent of document length, because an increased document length may increase the difference of paragraph and heading tags. However, it is not directly proportional to document length in the way that e.g. frequencies of verb groups are.

4. For many of the document structure features, such as the list of references and notes modules, an increased document length would increase the probability of more than one occurrence, but

probably not in a way that is directly proportional to the probability of other features depending on document length. Counts of these features have to be regarded on a scale similar to how counts of form elements are regarded.

Consequently the following principles for standardization have been chosen:

- The ratio of paragraph tags and heading tags is computed as in Equation 5.5 and then rescaled according to Equation 5.4.

$$PH^{count} = P^{count} - (H1^{count} + H2^{count} + H3^{count}),$$

$$x_{39} = \begin{cases} 0 & if\ PH^{count} = 0 \\ \log_2 PH^{count} & if PH^{count} \neq 0 \end{cases} \tag{5.5}$$

- For feature x_{40} (META elements), the raw counts are used and then rescaled according to Equation 5.4.

- Feature x_{38} (HTML Form elements) is treated on an ordinal scale, as given in Equation 5.6.

$$x_i = \begin{cases} 0 & if\ i^{count} = 0 \\ 0.5 & if\ i^{count} = 1 \\ 1.0 & if\ i^{count} > 1 \end{cases} \tag{5.6}$$

- The remaining markup features are treated as the linguistic and text-grammatical features are treated.

- The document structure features reflecting Tables, Figures, and Quotations are treated on an ordinal scale of six levels (Equation 5.7).

$$x_i = \begin{cases} 0 & if\ i^{count} = 0 \\ 0.2 & if\ i^{count} = 1 \\ 0.4 & if\ i^{count} = 2 \\ 0.6 & if\ i^{count} = 3 \\ 0.8 & if\ i^{count} = 4 \\ 1.0 & if\ i^{count} > 4 \end{cases} \tag{5.7}$$

- The document structure features reflecting Abstracts, Acknowledgements, Copyright statements, Definitions, Lists of references, Notes sections, Navigation bars, and Responsibility statements are treated on the same ordinal scale as feature x_{38}.

As can be observed, the possibility of differing variances is not accounted for in these standardizations. Nor are there any attempts to diminish the effect of outliers, which may have an unwanted effect on Equation 5.4.

5.6 Experimental questions

The experimental setup described in the previous sections has been designed for the purpose of examining the fourth question: **How do different definitions of genre spaces, classification models, and document features influence document genre classification?**.

The setup is somewhat restricted for an ideal situation in which the question could be fully answered. For instance, the pragmatic choice to ignore any computationally expensive tools for natural language preprocessing, necessarily makes the features derived too simple in comparison with what other research has shown to be successful. Moreover, the number of classification models and genre spaces possible to examine in full is subject to practical constraints.

A decision has been made to focus on a few genre spaces possible of covering the available corpus, and a few kinds of features that hitherto have not been investigated in full.

Considering the question of which classification model is best, there are some indications that SVM outperforms most other algorithms. With this in mind, together with the affordable amount of work, the question of algorithm evaluation with respect to comparisons with other algorithms is left aside. However, a few models have been chosen. The actual choices made are based on a) the author's familiarity with certain computational models, b) what is generally considered the best model, and, to a certain extent, c) what fits the kind of experiments to be performed.

Previous research and the investigations of how genres have been treated in the different disciplines reviewed in Part I of this work have been given interesting paths to follow in a more experimental fashion. Two kinds of features have been considered valuable to exploit further. These are the kind of features that may serve as indicators of speech acts, and the kind of features derived from logical document structures, or rather, the occurrences of certain document structure constituents (genre modules). The first experimental question is then:

1. Given a base set of features that are common in mainstream research, does adding features that represent document structure and speech act categories increase the classification performance?

Previous research into the classification along genre dimensions is to a large extent characterized by the employment of diverging sets of target genres. This is valuable as we need to know the influence of different genre spaces on the classification performance. However, much attention is also given to the feature selection and the choice of classification models. The variance of these two latter variables is therefore rarely consistently the same in different experiments. This makes comparisons of the different genre spaces employed quite unreliable. We may simply not be able to determine the impact of the choice of a genre space. Therefore, a second experimental question for this work is:

2. How do different genre spaces, with respect to granularities and the supposed nature of each genre, affect the classification performance?

With respect to the first question above, it follows that measures of classification performance must be determined both before and after the addition of sets of features. First, a general baseline for the data set and base feature set has to be set in order to make it possible to compare the performance estimators when new feature sets are introduced in the second step. With respect to the second question, the experiments will be divided into a set of initial and a set of secondary experiments. In the initial experiments of Chapter 6, the genre space is identical to the one employed for the original corpus annotation, which results in a set of mostly highly heterogeneous classes. In the secondary experiments of Chapter 7, the genre space for the *articles* class is adjusted to fit a more fine-grained notion of genre.

Chapter 6

Experiments with heterogeneous data sets

The *KI-04* corpus chosen has been used for a set of initial experiments which explore the extent to which the relatively simple base set of features ($\mathbf{x}_{base}$) fail to meet up with the performances attained in previous experiments — i.e. those performed by Santini (2007) and Meyer zu Eissen & Stein (2004). These initial experiments are intended to generate a kind of baseline, and are reported in Section 6.1. This chapter will thereafter report on the performance results with different classification models and adjustments to the corpus size, genre space, and minor adjustments to the base feature set. Finally, the impact of speech act features ($\mathbf{x}_{act}$) is preliminary assessed, by means of three measures for feature ranking introduced in Section 5.5.4.

In this and the succeeding chapter, the outcome of numerous experiments is reported. In order to make it a little bit easier to follow the trail, a few figures, that graphically illustrate the processes, have been included. In some cases, they will also report the performance figures otherwise found in the tables and in the normal flow of the text. Figure 6.1 illustrates the process of the initial experiments.

The first issue of the experiments in this chapter was thus to establish a kind of baseline to which the subsequent experiments could be compared. This baseline is the level of *accuracy* attained when

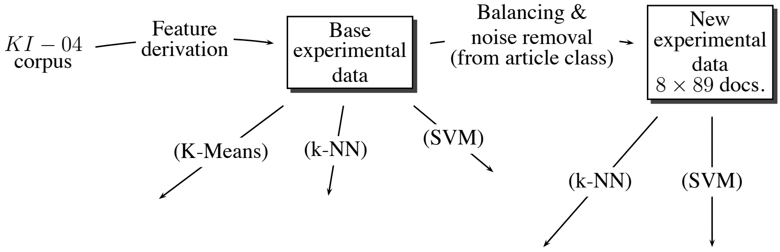

Figure 6.1: **Experimental configuration.**
The arrows directed downwards are labeled according to which classification model is employed in the evaluation of the respective data sets. The data sets are represented by shadowed boxes.

the base set of 39 features constituting the feature set used, as well as the *recall* and *precision* figures for classes of interest. For the establishment of this base level, the data set tried out was first assumed to be as similar as possible to the experiments by Santini (2007) and Meyer zu Eissen & Stein (2004), which means that the complete set of 1203 documents was used. The two documents that were infected by a virus were discarded. An additional baseline was established with a balanced subset of the source data, in order to maximize the entropy.

Two classification models were chosen for the experiments: the k-*NN* and the SVM algorithms as they are implemented in WEKA. Additionally, in some cases the *K*-means in its WEKA implementation was used. These models have been briefly described in the previous chapter.

For k-*NN* and SVM, a 10-fold cross-validation process was applied (see page 112 for this method). This validation process was in its turn repeated 10 or more times in order to estimate the variance of the results and subsequently to arrive at a confidence interval for the averaged figures. A confidence level of 95% was consistently used.[1] In some cases, output data for one of the resulting classifiers, from the

[1] In some cases, this repetition has been ignored, mostly because the figures are used for loose comparison only and the variances are expected to be reasonably low. When variance estimation has not been attempted, this is explicitly declared.

10-fold cross-validation, were reproduced from WEKA without any substantial change. What has been omitted are some measures which are not being discussed in this work. See Part I for an explanation of this output. For instance, confusion matrices are explained on page 110 and the following.

6.1 Baseline estimation

The results for the 10-fold cross-validation with k-*NN* and k set to 1 (i.e. 1-*NN*)[2] are in a first run as given in Figure 6.2, while the results for SVM are given in Figure 6.3. Here, the full *KI-04* set is used with the features extracted as described in Section 5.5. The parameter settings for SVM are the default settings and the same as Santini used in her experiments, except for the fact that logistic approximation is not used.

```
Correctly Classified Instances         516               42.8928 %
Incorrectly Classified Instances       687               57.1072 %
Total Number of Instances             1203

=== Detailed Accuracy By Class ===

Precision   Recall  F-Measure   Class
   0.622     0.543     0.58     articles
   0.586     0.535     0.56     discussion pages
   0.343     0.311     0.326    download pages
   0.417     0.432     0.424    help pages
   0.358     0.424     0.388    link lists
   0.363     0.38      0.371    non-private portrayals
   0.534     0.5       0.516    private portrayals
   0.367     0.365     0.366    shopping pages

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h   <-- classified as
 69  2  3 18 18  5  8  4 |  a = articles
  5 68  8 13 14  7  4  8 |  b = discussion pages
  4  5 47 15 23 19  6 32 |  c = download pages
  9 12  7 60 24  9 10  8 |  d = help pages
  7 11 18 16 86 28 14 23 |  e = link lists
  7  5 18  3 34 62  8 26 |  f = non-private portrayals
  7  4 10  9 16 13 63  4 |  g = private portrayals
  3  9 26 10 25 28  5 61 |  h = shopping pages
```

Figure 6.2: Results for one k-*NN* classification with k set to 1.

[2] Sometimes a shortened expression, e.g. 7-*NN*, is used, where the number represents the value of k.

```
Correctly Classified Instances         709               58.936 %
Incorrectly Classified Instances       494               41.064 %
Total Number of Instances             1203

=== Detailed Accuracy By Class ===

Precision   Recall  F-Measure   Class
   0.722     0.717     0.719    articles
   0.669     0.701     0.685    discussion pages
   0.448     0.49      0.468    download pages
   0.613     0.547     0.578    help pages
   0.619     0.616     0.617    link lists
   0.44      0.38      0.408    non-private portrayals
   0.702     0.73      0.716    private portrayals
   0.552     0.599     0.575    shopping pages

=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   <-- classified as
  91   5   4  11   3   3   9   1 |   a = articles
   2  89  10  11   4   4   4   3 |   b = discussion pages
   2   9  74   6  11  14   7  28 |   c = download pages
  13  13  12  76  12   6   2   5 |   d = help pages
   4   5  11   8 125  26   9  15 |   e = link lists
   4   3  28   7  27  62   4  28 |   f = non-private portrayals
   9   3   5   2  10   4  92   1 |   g = private portrayals
   1   6  21   3  10  22   4 100 |   h = shopping pages
```

Figure 6.3: Results for one SVM classification

These two figures only show the results for one of the 10 classifiers actually attained by randomizing the stratified folding. The averaged accuracy figures with error estimates are $42.2\% \pm 0.5$ for 1-*NN*, and $58.7\% \pm 0.4$ for SVM. Thus, SVM performs significantly better than 1-*NN*. This is no surprise, as many other experimental works show the same tendency of SVM to generally perform better than most other algorithms.

However, when k is set to 1, k-*NN* is usually considered to perform worse if class boundaries in the feature space are unclear. Increasing k may therefore also increase the performance. Table 6.1 shows only a slight increase in performance when k is increased (here, the results are given without performing 10 consecutive 10-fold cross-validations). The small increase of the accuracy figures when k is increased indicates that a large proportion of the misclassified instances is very distant in the feature space from what could be described as the center of their respective classes.

One may also draw the conlusion from the F-scores in figures 6.3 and 6.2, that instances of the classes *articles*, *discussion pages*, and

k										
1	3	5	7	9	11	13	15	17	19	21
42.9	44.6	44.9	46.1	46.3	47.5	48.2	47.7	48.4	48.0	48.0

Table 6.1: Accuracy figures with k-*NN* classification, where k varies from 1 to 21. 1203 documents and the 39 base features are used.

private portrayals are the easiest to predict, while *download pages*, *non-private portrayals*, *link lists*, and *shopping pages* often seem to be confused with one another. One may speculate in the way that this latter fact may be due to vague boundaries between the classes, as well as between genres, but such a statement needs to be supported by more thorough examinations. Comparing the figures with those of Santini (2007, p. 159ff) shows clear differences with respect to a worse F-score for *download pages* and *shopping pages*. This decrease may be due to the fact that Santini employed what she terms "genre-specific-word facets", a category of lexicals specifically tailored to the corpus at hand. It is not surprising that this in fact increases the recognition power, as tokens like 'FAQ', 'cart', 'download', 'credit card' etc. are likely to be discriminative for documents within the classes for which they are tailored. (Compare below, where the principle of Occam's razor is tested on a subset.)

The results for the SMO classification (i.e. the SVM algorithm implemented in WEKA) may be directly compared to the results reported by Santini (2007, p. 159) for her three different feature sets. This would indicate the suitability of the base set of features. Santini's results lie between a 62.5 and 70.2% accuracy. The results here are significantly worse, but that is no surprise, given the more simplified feature set used. What was expected was actually a much worse result.

Finally, an interesting question is what the supervised learning approach contributes to the result, compared to an unsupervised approach. If the documents are disattached from any assumptions on class or genre adherence, their mere features are used for clustering into eight clusters. Consequently, if a perfect match could be attained when the resulting cluster assignments are matched against the gold standard, the feature set in itself would be sufficient for a classification and the expensive task of a gold standard annotation would be unnecessary. This would in fact imply that there is a very simple relationship between artefactual features and genre document classes, which would be highly surprising.

K-means clustering is therefore applied with 10 different numbers of random initial seeds (between 8 and 26).[3] The result in terms of an overall accuracy is as low as $30.8\% \pm 0.8$.[4] This may indicate several things. For instance, clustering may be considered inappropriate for this kind of task, as artefact variation is a common feature within the genres decided upon for the collection.[5] A more fine-grained genre space would then be preferred, if clustering were to be fruitful. However, evaluating such a task would probably require a much larger corpus and a thorough reannotation.

6.2 The principle of Occam's Razor

Before moving on, it is appropriate to regard the principle that states that simpler models are always to be preferred. Santini included a few, intuitively chosen, genre-specific words in her feature sets. This is perfectly plausible for some of the original *KI-04* classes. In order to take this idea a bit further, the *articles* class is removed and the *private* and *non-private portrayals* are collapsed into one class, so that we have six target classes instead of eight. The *articles* class is removed because intuitively, it is difficult to find a set of words that would be considered discriminative for that class, and the two *portray-*

[3] In *WEKA* there are no means to control the way seeds are chosen, which otherwise would have been valuable to examine further. Choosing different prototypical documents as 8 initial seeds, for instance, may have yielded better performance, but would also imply a semi-unsupervised approach. In addition, the documentation of *WEKA* does not specify how the implementation decides on reduction principles when the number of seeds exceeds the number of target clusters.

[4] This confidence interval may be unreliable, because there is no obvious justification for the assumption that the accuracy of clustering when changing the number of seeds is actually normally distributed.

[5] Cf. how the instances of the *discussion* class demonstrate occurrences of three different document structures, discussed in Section 5.3.2.

als-classes are collapsed as the words that come to mind, or that are observed for such pages, will most surely be frequent in both classes. The features created are six, consisting of the counts of occurrences of each category of word tokens considered specific for the respective class. Ocular inspections of random samples from each class have been guided the choice of specific tokens.

The categories of words chosen for the six features are given in Table 6.2. Occurrences of the words in the full text are not considered, as only the titles of the documents are processed.[6]

Help pages	help, faq, questions, answers
Portrayals	welcome, home, homepage
Link lists	links, directory, list, link, linklist, references
Shopping pages	store, shop, shopping
Download pages	download, downloads, free
Discussion pages	discussion, forum, discussions

Table 6.2: The six hand-tailored features of word tokens.

The results for one of the classifiers are presented in Figure 6.4, where the variance after 10 consecutive 10-fold cross-validations is almost zero. The figures are therefore more than indicative. This also indicates that the number of features affects the reliability of the performance figures. Fewer features seem to decrease the estimated error. As can be seen from the confusion matrix, the portrayals confuse the algorithm. Far too many instances are classified as portrayals, probably due to the fact that none of the words of the adequate feature occur in the title. Precision is remarkably high for all other classes. Similar, but slightly worse performance figures, are attained with k-NN and K-means, but not reported here.

However, in order to estimate whether it is the 6 features that give these results, or the modification of the genre space of the *KI-04* corpus, we have to compare the results with the base feature set as well. These are given in Figure 6.5, and even though the variance has not

[6] This way of ignoring the full text is probably more conformant with how humans mostly approach the task of genre identification when first exposed to a previously unseen document.

```
Correctly Classified Instances         786              73.0483 %
Incorrectly Classified Instances       290              26.9517 %
Total Number of Instances             1076

=== Detailed Accuracy By Class ===

Precision   Recall  F-Measure   Class
   0.95      0.598     0.734    discussion pages
   0.957     0.742     0.836    download pages
   0.948     0.791     0.863    help pages
   0.983     0.576     0.727    link lists
   0.51      0.979     0.671    portrayals
   0.989     0.527     0.688    shopping pages

=== Confusion Matrix ===
   a   b   c   d   e   f   <-- classified as
  76   0   2   1  48   0 |   a = discussion pages
   0 112   2   0  37   0 |   b = download pages
   1   1 110   0  27   0 |   c = help pages
   0   2   1 117  82   1 |   d = link lists
   3   1   1   1 283   0 |   e = portrayals
   0   1   0   0  78  88 |   f = shopping pages
```

Figure 6.4: Results for one SVM classification, 6 classes, 6 word token features.

been estimated, it is clear that the better performance gained by the 6 word token features is not a result of a changed genre space. This is somewhat surprising, and one could conclude that designing algorithms based on genre specific lexicals should be preferred to more expensive feature sets — at least for some genre spaces. However, one cannot rely on significant titles being always given and on all valuable genre classes exhibiting a predictable and significant lexical repertoire.

6.3 Balancing and purifying the data set

There are two problems with the data set used this far. First, the class distributions are somewhat skewed, which may affect the results to some extent. This is probably the reason why Meyer zu Eissen & Stein (2004) used only a subset of 8 × 100 documents in their experiments. Second, the ocular inspection reported in the previous chapter showed that the *articles* class was very noisy, with a few obvious misclassifications and a few documents in German. Therefore, the experiments of Section 6.1 were repeated with a smaller and balanced data set,

```
Correctly Classified Instances         635               59.0149 %
Incorrectly Classified Instances       441               40.9851 %
Total Number of Instances             1076

=== Detailed Accuracy By Class ===

 Precision   Recall  F-Measure   Class
0.649     0.669     0.659    discussion pages
0.487     0.483     0.485    download pages
0.667     0.547     0.601    help pages
0.613     0.586     0.599    link lists
0.599     0.616     0.608    portrayals
0.547     0.623     0.583    shopping pages

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
  85   8  13   4  13   4 |   a = discussion pages
   7  73   7  15  24  25 |   b = download pages
  15   8  76  15  18   7 |   c = help pages
   8  12   8 119  40  16 |   d = link lists
  12  27   9  29 178  34 |   e = portrayals
   4  22   1  12  24 104 |   f = shopping pages
```

Figure 6.5: Results for one SVM classification, 6 classes, base features.

consisting of 89 documents from each class — this will be referred to as the 8 × 89 data set in the following.[7] Removing misclassifications and the documents in German is expected to increase the performance. K-means was not run.

Actually, the results given in Figure 6.6 illustrate an accuracy of $60.5\% \pm 0.5$, which is just a small, although significant, increase in performance, while k-*NN* shows different effects in performance (41.5%, with k =1, and $52.2\% \pm 0.4$ with k = 13, distance weighting active).

If the confusion matrix is considered, one may make a few interesting observations in pair-wise confusions (where the *articles* class is part of the pair). There is no confusion whatsoever between *articles* and *shopping pages*, and a very limited confusion between *articles* and *discussion pages*. Contrariwise, a considerable confusion occurs between the *articles* and *help pages* classes.

Figure 6.1, that contained part of the experimental configuration, can now be completed with the values for the output in Figure 6.7.

[7] 89 is the number of documents in the *articles* class that were considered appropriate to assign to the *articles* class. See Section 5.3.2

```
Correctly Classified Instances         428               60.1124 %
Incorrectly Classified Instances       284               39.8876 %
Total Number of Instances              712

=== Detailed Accuracy By Class ===

Precision   Recall  F-Measure   Class
   0.708     0.708     0.708    articles
   0.643     0.708     0.674    discussion pages
   0.529     0.517     0.523    download pages
   0.563     0.506     0.533    help pages
   0.607     0.573     0.59     link lists
   0.477     0.472     0.475    non-private portrayals
   0.72      0.753     0.736    private portrayals
   0.548     0.573     0.56     shopping pages


=== Confusion Matrix ===

  a  b  c  d  e  f  g  h   <-- classified as
 63  2  4 11  2  2  5  0 |  a = articles
  0 63  3  9  1  4  4  5 |  b = discussion pages
  2  9 46  3  4  8  2 15 |  c = download pages
 11 10  5 45  5  5  5  3 |  d = help pages
  3  5  4  3 51 11  7  5 |  e = link lists
  3  3  9  6 12 42  1 13 |  f = non-private portrayals
  7  4  2  2  5  1 67  1 |  g = private portrayals
  0  2 14  1  4 15  2 51 |  h = shopping pages
```

Figure 6.6: Results for SVM, when the data set is balanced and "purified".

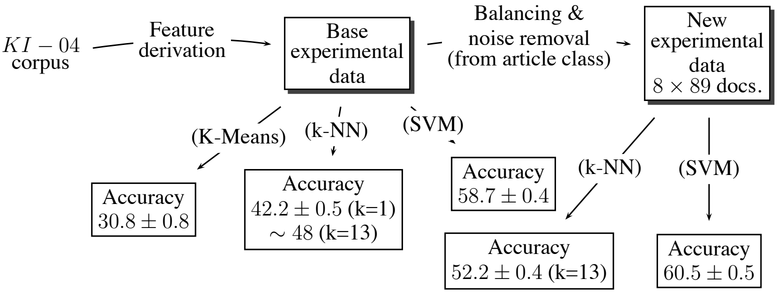
$KI-04$ corpus
Feature derivation
Base experimental data
Balancing & noise removal (from article class)
New experimental data 8 × 89 docs.
(K-Means)
(k-NN)
(SVM)
Accuracy 30.8 ± 0.8
Accuracy 42.2 ± 0.5 (k=1) ∼ 48 (k=13)
Accuracy 58.7 ± 0.4
(k-NN)
(SVM)
Accuracy 52.2 ± 0.4 (k=13)
Accuracy 60.5 ± 0.5

Figure 6.7: Experimental configuration from Figure 6.1 with results.

The amount of confusion with respect to the discrimination between *articles* and *discussion pages*, and between *articles* and *help pages* are interesting. Therefore, two more data subsets are constructed with these pairs, to see whether these relationships hold true even when only two classes are considered.

```
Correctly Classified Instances         166               93.2584 %
Incorrectly Classified Instances        12                6.7416 %
Total Number of Instances              178

=== Detailed Accuracy By Class ===

Precision   Recall  F-Measure   Class
   0.943     0.921     0.932    articles
   0.923     0.944     0.933    discussion pages

=== Confusion Matrix ===

  a  b   <-- classified as
 82  7 |  a = ar
  5 84 |  b = di
```

Figure 6.8: Results for one SVM classification, with two classes, *articles* and *discussion pages*.

Figure 6.8 shows the results for one SVM classifier operating on the *articles* versus *discussion pages* pair. When running SVM with 10 different stratifications of the 10-fold cross-validation, the accuracy is 92.3% ± 0.6. However, running the same experiment with simple *K*-means shows that clustering is strangely enough as accurate, with an overall accuracy of 93.3%.[8] We know that *discussion pages* are most often constitued by being either indices to discussion postings, a set of discussion postings, or one discussion posting. It seems that these artefactual forms can be discriminated fairly easy from the heterogeneous class of *articles*, with almost no human intervention at all.

Figure 6.9 shows the results for discriminating *articles* from *help pages*, in which 79.6% ± 0.7 accuracy is attained, if performed over 10 different cross-validations. For this task, clustering is much more inaccurate and gives an accuracy of only 57.9%.

The results in figures 6.9 and 6.8 show (not surprisingly) that de-

[8] It has to be noted that *K*-means seems to be extremely sensible to the number of seeds. Almost invariably exactly 93.3% is attained, but for high numbers of seeds the error may be doubled.

```
Correctly Classified Instances         145               81.4607 %
Incorrectly Classified Instances        33               18.5393 %
Total Number of Instances              178

=== Detailed Accuracy By Class ===

   Precision   Recall  F-Measure   Class
      0.833     0.787     0.809    articles
      0.798     0.843     0.82     help pages

=== Confusion Matrix ===

  a  b   <-- classified as
 70 19 |  a = articles
 14 75 |  b = help pages
```

Figure 6.9: Results for SVM with 2 classes, *articles* and *help pages*

creasing the number of classes to consider for a classifier generally increases its capability to recognize the classes. What then, if we ask to what extent the same model manages to discriminate between documents in the *articles* class from the rest? In this case, accuracy is much less interesting, since the number of *non-articles* are far too many compared to the *articles*. What matters here is only the precision and recall for the *articles* class. It would otherwise be simple to attain 87.5% accuracy, just by assigning each document to the *non-articles* class. The 8×89 subset was relabeled, so that members of the *articles* class are given the label `pos` and the remaining 7×89 documents the label `neg`.

Here, the precision for the *articles* class is $83.6\% \pm 0.9$ and the recall only $61.2\% \pm 1.0$, with an F-score of $70.7\% \pm 0.6$, applied on the 8×89 subset. The precision is remarkably good. One of the resulting classifiers misclassifies 9 documents as *articles*, but fails to recognize as many as 37 *articles*. 6 of the *non-articles* are *help pages*, and one each belongs to the classes *download pages*, *private portrayals*, and *non-private portrayals*.

The misclassified *download page* is a verbose and fairly instructive page that offers the possibility to download material from the "CIA world fact book". The misclassified *non-private portrayal* is a "frontpage"[9] for "The Magma Computational Algebra System", offer-

[9] With frontpage is denoted a document that mainly delivers links to other related documents and thus functions as a pathway into resources offered by the provider that

ing links to closely related pages. The misclassified *private portrayal* is a publication list and a verbose account of the research interests of one James F. Allen. None of these three documents can be considered typical for the classes which they have been assigned to by the compilers of *KI-04*.

This corroborates to some extent the assumption that *help pages* in general constitute the class that is the most similar to *articles*. *Help pages* are often instructive and thus linguistically similar to the tutorials and how-to's of the *article* class. In the *KI-04* corpus *help pages* are also often topically similar to *articles*, while the other classes often treat topics that do not occur in the *articles*. This may indicate that the topic is influential in genre classification, and should therefore not be straightforwardly considered orthogonal to genre adherence.

The experiments this far have established a few base levels of performance that can be used to compare the succeeding experiments. However, in order to assess that some features do not introduce unwanted noise, there is also a possibility to examine the features in themselves more thoroughly, according to different measures that measure their impact on classification tasks.

WEKA implements, among many measures, *Information Gain*, *Gain Ratio*, and χ^2 as procedures that produce lists of all features ranked according to their estimated significance. Table 6.3 contains the ranked results of these procedures. As we are mainly interested in how the features discriminate *articles*, the results are derived from the 8×89 data set where *articles* are discriminated from the rest. The rankings thus reflect their power for discriminating between *articles* and *non-articles*.

Information Gain	*Gain Ratio*	χ^2
anchors	suc	anchors
suc	anchors	suc
resc	resc	resc
doc_lengthT	eqc	eqc
img	reic	doc_lengthT

Continued on the next page

set up the frontpage.

Continued from the previous page

table	apc	indefinite_pronouns
ph_relation	pre	conc
ampl	conc	ph_relation
dwntone	ph_relation	ampl
conc	table	dwntone
eqc	dwntone	img
indefinite_pronoun	anc	table
temp_adv	enc	reic
suasives	ampl	temp_adv
priv_vb	doc_lengthT	anc
pronouns_2nd	priv_vb	apc
excl_marks	img	suasives
commas	excl_marks	pronouns_2nd
reic	pronouns_2nd	enc
anc	indefinite_pronouns	priv_vb
pronouns_1st_pl	commas	commas
apc	stops	excl_marks
form	pronouns_1st_pl	pre
colons	suasives	form
enc	qu_marks	pronouns_1st_pl
stops	colons	spat_adv
li	meta	colons
spat_adv	temp_adv	inc
qu_marks	li	li
pre	form	pub_vb
pub_vb	inc	stops
inc	spat_adv	qu_marks
meta	pub_vb	meta
token_length	token_length	token_length
disc	quot_marks	pronouns_1st_sing
refc	tec	refc
pronouns_1st_sing	disc	tec
quot_marks	refc	disc
tec	pronouns_1st_sing	quot_marks

Table 6.3: Feature rankings based on three measures and the 8 × 89 subset of *articles* and *non-articles*.

When the ranking is considered from the bottom and up, it can be observed that in all three cases the six lowest ranked features are all the same. If this ranking is to be relied upon, we should be able to train and test the algorithm with these features removed, without any expected loss in performance. Then, the only difference is actually that one of the *articles* not recognized before is now recognized. It really seems that these six features are more or less irrelevant, but as it cannot be determined that they do any significant harm, they will be retained in the following experiments.

Now, we take the opposite approach and include only the five highest ranking features, on the premise that a simpler model is always to be preferred. The precision decreases to 72.7% (previously 83.6%) and the recall to 53.9% (previously 61.2%) for one of the resulting classifiers (no confidence interval determined). As a last experiment, only the ratio of anchors (the most effective feature) is used, which yields the catastrophical recall of 15.7% and a precision of 63.6%.

Figure 6.10 is an illustration of the experiments performed in this section with some of the results given.

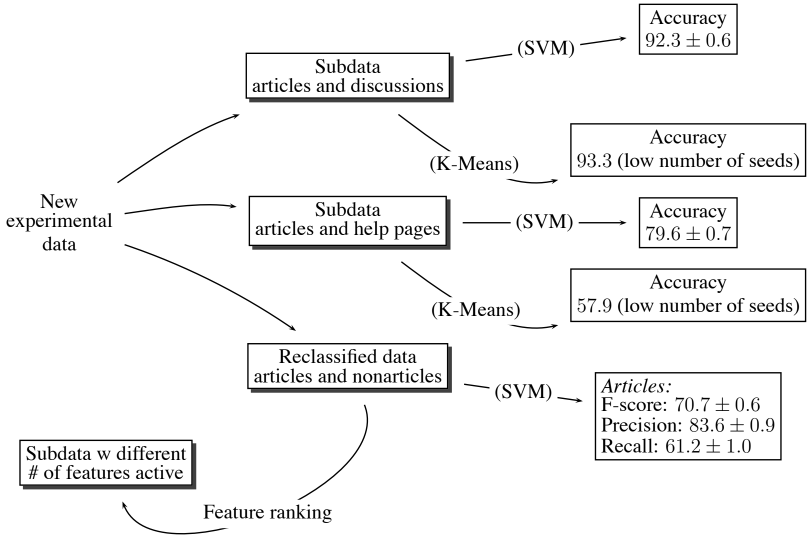

Figure 6.10: Experimental configuration for testing subsets of the data

All these figures indicate that a high recall for *articles* is more difficult to accomplish than an acceptable precision. No attempt has been made to identify the kinds of *articles* the classifiers generally fail to recognize.

In the next chapter, the investigation of features will be taken up again, this time with respect to the two kinds of features of particular interest for this work (i.e. speech-act features and document structure features). Some levels of performances acquired in this chapter will be used to assess the impact of the speech act features, while the document structure features require a somewhat different approach, since we do not have access to such features for the entire data set.

However, the remaining 89 *articles* that are the only ones with identified document structure features, can be suspected to be a too small subset. Therefore, this subset has been extended by 54 new documents, and we need to see how this affects the baseline levels acquired this far.

6.4 Extending the *articles* class

Two initial experiments are performed after the addition of the 54 new documents, on the premise of the experiments that were performed before the additions. As the primary interest lies in the *articles* class, the question is if a purification and extension of this class would increase performance with respect to just the *articles* class, compared to the experiments reported briefly on page 198. The case of 8 classes is deemed less interesting and only the binary classification problem is thus evaluated.

First, representations of the 54 documents are added to the 8×89 subset, and SVM is applied. Then the same addition is made to the entire `KI-04` dataset. The precision, recall and F-scores for *articles* attained are given in row two and three of Table 6.4, and a sample confusion matrix in Figure 6.11.

The results are significantly different from before the introduction of the 54 documents. The increased performance may be partly explained by the less arbitrary compilation of the new *articles* subset, as opposed to the heterogeneity of the previous *articles* subset, and the increased ratio of *articles* versus the other classes.

Subset	Precision	Recall	F-score
712	83.6 ± 0.9	61.2 ± 1.0	70.7 ± 0.6
712 + 54	87.4 ± 1.0	72.4 ± 0.5	79.2 ± 0.5
1165 + 54	89.4 ± 0.5	70.6 ± 0.5	78.9 ± 0.3

Table 6.4: Results for classification with the extended subsets. The first row gives the previous results for classification without the 54 new *articles*.

```
=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.741     0.024      0.876     0.741     0.803    pos
  0.976     0.259      0.943     0.976     0.959    neg

=== Confusion Matrix ===

   a   b   <-- classified as
 106  37 |   a = pos
  15 608 |   b = neg
```

Figure 6.11: Sample confusion matrix for SVM classification with the extended *articles* class and the 8×89 subset, two classes.

What is interesting when comparing one of these classifiers with one of those from the previous 8×89 subset is that the same number of *articles* is unrecognized, compared to what was the case before the introduction of the 54 new documents. This means that the 54 new *articles* are probably all easily recognized (if the unrecognized documents are the same as before). The conclusion may be that since adding these *articles* does not affect the number of unrecognized *articles*, the 54 new documents are more or less typical for the *articles* class.

The same comparison, however, shows that the algorithm has included 6 more *non-articles* in the *articles* class. This may be explained as dependent on the fact that some of the 54 *articles* are similar to some *non-articles* in a way that confuses the algorithm with respect to these *non-articles*.

On the contrary, when the second, larger, subset is used, the number of *non-articles* included is slightly decreased, while the number of *articles* not recognized is slighly increased. This may indicate that some of the 623 *non-articles* are very noisy with respect to the features.

6.5 Adding speech act features

With the largest set above in mind, it is now possible to add the speech act features. The results for one of the resulting classifiers are given in Figure 6.12. The averaged precision, recall and F-score for *articles* are $89.8\% \pm 0.6$, $73.7\% \pm 0.4$ and $81.0\% \pm 0.4$, respectively. We can see that there is no significant difference in terms of precision, even though there is a slight significant increase of recall. The results for these speech act features are thus somewhat promising for future experiments.

6.6 Summary of the initial experiments

This chapter has arrived at some conclusions, indications and performance figures of importance for further experiments. It has been shown that

```
Precision   Recall  F-Measure   Class
   0.882     0.734     0.802    pos
   0.965     0.987     0.976    neg

=== Confusion Matrix ===

    a    b   <-- classified as
  105   38 |    a = pos
   14 1062 |    b = neg
```

Figure 6.12: Results for one SVM-classifier with 2 classes, *articles* and *non-articles*, speech act features added

- The SVM algorithm is generally superior to k-*NN* when large data sets are to be classified. Unsupervised classification (in the form of *K*-means) should probably not be relied upon in such cases, except for when discriminating between very dissimilar classes.

- The base set of features performs only slightly worse than feature sets used in previous research on similar document collections.

- The set of features in combination with SVM and its default settings in WEKA seem to be relatively well suited for the recognition of documents in the *articles* class.

- Adding and subtracting small numbers of features do not change the performance to any greater degree.

- For *articles*, the recall is generally lower than the precision, or, in other words, excluding *non-articles* is easier than including *articles*.

- Decreasing the number of classes radically improves the performance.

- Speech act features show some small indications on improving classification performance in terms of recall.

Table 6.5 summarizes the results, where π, ρ and F-score always refer to performances with respect to the *articles* class. The figures given are in percentages. Where no confidence interval is given, the values should be regarded only as indicative, and therefore they are not given as real-valued scores.

Data Set	Classes	Algorithm	Accuracy	π	ρ	F-score
Full *KI-04* set	8	SVM	58.7 ± 0.4	72	72	72
		k-NN	42.2 ± 0.5	62	54	58
		K-means	31			
8×89		SVM	60.5 ± 0.5	71	71	71
	ar/di	SVM	92.3 ± 0.6	94	92	93
	ar/di	K-means	93.3			
	ar/he	SVM	79.6 ± 0.7	83	79	81
	ar/he	K-means	57.9			
	ar/non-ar	SVM		83.6 ± 0.9	61.2 ± 1.0	70.7 ± 0.6
Extended Full *KI-04* set	ar/non-ar	SVM		89.4 ± 0.5	70.6 ± 0.5	78.9 ± 0.3
Extended Full *KI-04* set with $\mathbf{x}_{act}$	ar/non-ar	SVM		89.8 ± 0.6	73.7 ± 0.4	81.0 ± 0.4
Extended 8×89	ar/non-ar	SVM		87.4 ± 1.0	72.4 ± 0.5	79.2 ± 0.5

Table 6.5: Summary of performance measures for the large data sets. The “Extended” sets comprises the 54 documents not present in the *KI-04*.
$\mathbf{x}_{act}$ is the set of features representing speech acts.
π = precision for *articles*,
ρ = recall for *articles*.

Chapter 7

Experiments with the *articles* class

The class of *articles* has been annotated in two more ways than the full *KI-04* corpus. This reannotation has been described in Section 5.3.2. First, the number of occurrences of the genre modules (document structure features) in each document have been recorded in a comment section of the documents, then each document has been mapped to a class in another more fine-grained classification scheme. These annotations are algorithmically added to the feature representations of each document. In other words, the class of *articles* has been reannotated and a new representation with a higher feature dimension has been produced for each instance in the subset of the source class.

A considerable drawback here is that we, unfortunately, only have $89 + 54$ instance of the *articles* class remaining. Variance figures will indicate the significance of this drawback, but they could be expected to be higher than for the experiments of the previous chapter. The long table starting at page 147 lists the fine-grained classes of genre artefacts identified and included in the original *KI-04* corpus. If only the 89 remaining documents are considered, there are twelve classes of which two consists of only one document each. This would make a 10-fold stratified cross-validation completely unreliable, since there cannot be more folds than there are documents in the smallest class. In addition, an algorithm needs to have a considerable amount of documents for each class to learn from. The minimum level of class members necessary for an effective training to take place cannot be known beforehand, but it seems reasonable to think in a direction where the number of documents in the corpus is higher than the number of features, and the number of training documents for each class is at least 30.

This is why the following experiments are applied on a grouping of the source classes into three generalized classes, compiled according to the principles laid out in Section 5.3.2, from page 143 and onwards. The fourth group of “nonconsistent *unfit* documents” has been ignored. They will be referred to as *research articles*, *tutorials* and *reports*, respectively. The distribution of the classes is fairly even, 46, 49, and 48 instances each.

7.1 Validating the genre space

The previous experiments have all been performed with the same target classes for which the *KI-04* corpus was compiled. The validity of this previous genre space has been questioned and discussed, but for benchmarking reasons left unchanged. In the experiments of this chapter, benchmarking issues are no longer at stake and a completely different genre space is therefore introduced. This motivates a tentative investigation of the validity of this space, not the least because its task complexity, with respect to a human mind, must be assessed to some degree. If users radically disagree on the correct class assignments, any attempt to evaluate classification based on human class assignment will be debatable.

In order to validate the genre space, a small user study was undertaken for a subset of 30 documents from the source *articles* class of the *KI-04* corpus and the extended set of 54 documents. Care was taken to get a subset that was balanced with respect to the three compound classes of argumentative, descriptive and instructive texts, as well as including four of the documents that were judged as unfit for the *articles* class.

After given two lectures on the topic of non-fictional genres and genre classification, approximately 20 LIS undergraduate students were divided into four groups. The four groups were given the 30 documents and told to cooperate within the group, in subdividing these documents into 4-8 classes corresponding to their estimated genre adherence.[1] When they had done this and the group felt confident with the class assignments, they were asked to find a description and a label for the classes. They were explicitly told not to communicate with members from any other group.

The four groups came up with 6, 7, 7, and 8 different classes, respectively. In total, the groups suggested 23 different labels, of which several were rather similar but not identical. There were only four labels that, in exact terms, occurred in the assignments of more than one group — `scientific articles`, `portals`, `glossaries`, and `reports`. In the case of `scientific articles`, the only label assigned by three groups, there was no perfect agreement between the three groups on the members of this class. However, five documents were agreed upon as being `scientific articles` by all of these three groups. Two documents were agreed upon as being `reports` by the two groups who came up with that label.

The group that did not suggest the label `scientific articles` instead used a more generic label, `article`, for the five documents that were agreed upon as being `scientific articles` by the other groups. All three groups used labels that can easily be interpreted as variations on labels for instructive texts (textbook-like) introducing the management of some technology. These five documents can be considered prototypical.

If we take scientific articles as being synonymous with *articles*, and the three labels on instructive material as being synonymous, we have a perfect agreement on 9 (30%) of the 30 documents. One particularly difficult document to classify was the ”home page for the axiom of choice – an introduction and links collection” (article_5364813528), which was classified as *infotext*, *personal reflection*, *overview*, and *portals* by the four groups. This document is a

[1] The lower limit was set in order to reflect the number of columns of the table at page 147.

fairly formal presentation of a mathematical theory, ending with a list of links to related material. It is obviously difficult to classify as long as disjoint classification is assumed.

The most striking general observation that can be made on this very restricted investigation is that, as in the case study presented in Section 3.6.1 on the WebGenreWiki (2008), variation in the choice of labels is very high. In addition, a perfect agreement on both labels and class assignment seems to be extremely rare. However, one must take account of the fact that if a set of labels had been predefined and given to the students, the classification result in terms of agreement may have been much better. For instance, the difference between `project history` and `project description`, or between `report` and `report (project summary)`, may be resolved with a predefined list of class labels.

It is promising to note that for the 9 documents for which perfect agreement occurred the assignments conformed with the reannotation assignments of the corpus, and that for several others the labels may be interpreted as signifying a category that is also conformant with the reannotation assignments. The most obvious confusion occurs when the documents include an extensive section of links to related material and may therefore be interpreted as a combination of predominantly argumentative, descriptive or instructive texts, and link collections.

So, conclusion, the task is fairly complex, but, judging from this subset of 30 documents, there does not seem to be a too large confusion around the main part of the subcollection. The next section will thus proceed by presenting the results of the experiments targeting the three compound classes with base features only.

7.2 Baseline estimations

The experiments with the *articles* class will be performed in order to answer the question of what impact the features of document structure and speech act verbs have on the classification performance. Figure 7.1 illustrates how the experiments are configured for these objectives.

It was decided to try both SVM and k-*NN* classification as a start-

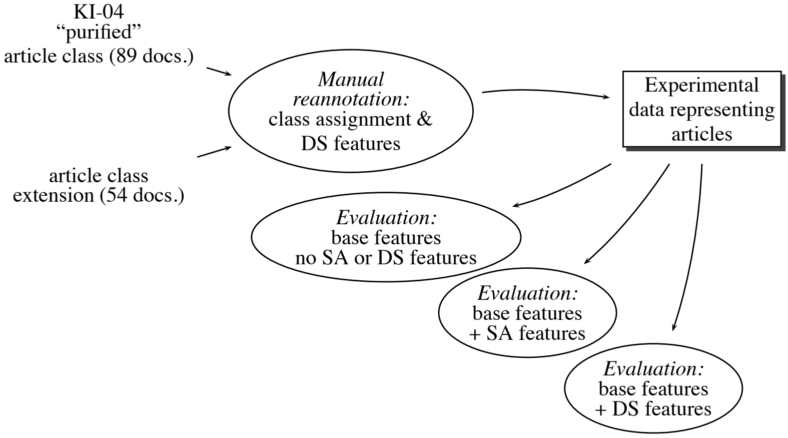

Figure 7.1: Experiments with the *articles* class

ing point, to determine a tentative baseline. The results for one classifier of the SVM and 7-*NN* types respectively are shown in figures 7.2 and 7.3. Different values of k were tentatively tried out, but no large differences were discerned, so the value of 7 is fairly arbitrary.

Table 7.1 shows the results with confidence intervals and, quite surprisingly, there is no significant difference between the two algorithms when accuracy is considered.[2] However, the SVM algorithm is clearly better in terms of recognizing the instructive class (*tutorials*) and the 7-*NN* algorithm in recognizing the class of *reports*. Generally, *research articles* are easier to recognize.

Running K-means yields accuracy above 55% if 5, 6, or 7 arbitrary initial seeds are given, but varies considerably depending on the number of initial seeds. This is interesting as the difference between a clustering approach and a classification approach (in terms of accuracy) is here considerably smaller than in the experiments in the

[2] It should be noted that the number of runs for the variance estimation has been increased from 10 to 15 here, as the variance is expected to be higher.

```
Correctly Classified Instances          87               60.8392 %
Incorrectly Classified Instances        56               39.1608 %
Total Number of Instances              143

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.653     0.202      0.627     0.653     0.64     tutorials
  0.479     0.221      0.523     0.479     0.5      reports
  0.696     0.165      0.667     0.696     0.681    research articles

=== Confusion Matrix ===

  a  b  c   <-- classified as
 32 10  7 |  a = tutorials
 16 23  9 |  b = reports
  3 11 32 |  c = research articles
```

Figure 7.2: Results for one SVM-classifier with the 3 *articles*-classes as target classes, only the 39 base features used.

```
Correctly Classified Instances          90               62.9371 %
Incorrectly Classified Instances        53               37.0629 %
Total Number of Instances              143

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.49      0.085      0.75      0.49      0.593    tutorials
  0.604     0.179      0.63      0.604     0.617    reports
  0.804     0.289      0.569     0.804     0.667    research articles

=== Confusion Matrix ===

  a  b  c   <-- classified as
 24 11 14 |  a = tutorials
  5 29 14 |  b = reports
  3  6 37 |  c = research articles
```

Figure 7.3: Results for one 7-NN-classifier with the 3 *articles*-classes as target classes, only the 39 base features used.

		F-scores		
Algorithm	Accuracy	Tutorials	Reports	Research articles
SVM	61.7 ± 0.9	64.7 ± 0.9	55.8 ± 1.9	65.1 ± 1.0
7-NN	62.1 ± 0.9	54.7 ± 1.5	64.0 ± 1.2	66.8 ± 1.2

Table 7.1: Results for SVM and 7-NN classification, the three-class problem.

previous chapter. Even though the size of the data here is smaller and probably less noisy, further investigations of clustering on the basis of this kind of genre spaces could be fruitful.

When evaluating the features on the basis of the three measures that were also used in the previous chapter (and presented in Section 5.5.4), it turns out that only nine features, listed in Table 7.2, score above zero impact.

list items
private verbs
document length
relation between paragraph and heading tags
downtoners
concessive connectives
average token length
indefinite pronouns
public verbs

Table 7.2: Features effective for the *articles* subset.

This feature evaluation confirms what Kim & Ross (2008) have already stated, that the optimal feature set is highly genre-dependent. In the previous chapter where the feature ranking was calculated on the basis of *articles* and *non-articles*, several categories of connectives were among the best ranked features, as well as `anchor tags`. These are now completely ineffective, if these measures are to be trusted. Contrariwise, *list items* and *public verbs* were ranked low earlier but are now considered effective.

Reducing the features to only these 9 features yields an accuracy of $60.9\% \pm 0.7$ for SVM, whereas the F-scores are 57.8 ± 1.1 (*tutorials*), $63.2\% \pm 0.8$ (*reports*), and $62.6\% \pm 1.2$ (*research articles*). Thus, there is no significant change in accuracy, even though there is when the three classes are considered alone. It is often stated that one should choose the simplest algorithm, and one could therefore conclude that using only the 9 features is the best choice, if only the overall accuracy matters. However, in the following the full set of features is used.

7.3 Adding speech act features

Adding the 5 speech act verb categories actually decreases the results for SVM into an accuracy of only $59.2\% \pm 1.0$, while the F-scores are $61.0\% \pm 1.3$ (*tutorials*), $51.8\% \pm 1.3$ (*reports*), and $61.5\% \pm 1.4$ (*research articles*).

One could thereby conclude that speech act features actually add confusion when it comes to discriminating between these kinds of classes. However, one more investigation is worth considering, namely to evaluate the significance of the 5 speech act features as before.

Feature evaluation shows that only `commissives` and `verdictives` have anything to contribute. Removing the other three features yields an accuracy of of $59.0\% \pm 1.0$, while the F-scores are $62.4\% \pm 1.4$ (*tutorials*), $52.1\% \pm 1.1$ (*reports*), and $62.1\% \pm 1.6$ (*research articles*). There is actually no significant difference at all. Running K-means on this last configuration yields an accuracy above 60% if K-means is given 5 initial seeds, but as before it may yield around 50% for other numbers of initial seeds. A cumbersome but interesting investigation would be to study the agreement between classifications and cluster assignments here, because if this agreement is high, one has to thoughtfully investigate what the reasons are for this. However, this is left out for further investigation.

7.4 Adding document structure features

Now, since the preceding section indicated that speech act categories may be confusing with respect to our three-class problem, these features are left out in the following experiments, where focus is on the document structure features. If we start by adding all document structure features, we get an overall accuracy of $62.7\% \pm 1.3$, while the F-scores are $64.8\% \pm 1.9$ (*tutorials*), $56.7\% \pm 1.7$ (*reports*), and $65.9\% \pm 1.4$ (*research articles*).

These results show no clear significant differences with respect to the $\mathbf{x}_{base}$ set, but they have possible positive indications, as all averages are above the previous results.

Feature evaluation by means of the aforementioned three ranking measures, indicates that only three of the document structure features have anything at all to contribute to the classification, namely `reference lists`, `figures`, and `abstracts`. Consequently, SVM was run with only these features added, with a resulting overall accuracy of $62.8\% \pm 1.1$, while the F-scores are $65.2\% \pm 1.2$ (*tutorials*), $57.8\% \pm 1.4$ (*reports*), and $65.9\% \pm 1.4$ (*research articles*). Again the results are slightly improved, but without any significance.

When running K-means, accuracy is again almost as good as classification. This also holds when several different numbers of initial seeds are given. Figure 7.4 shows the confusion matrices for the stratified 10-fold classification with the best results (65% accuracy) versus the best cluster assignments (60.8% accuracy). One may discern that the most obvious differences concern the confusion between *tutorials* and *reports*, in that many more *tutorials* have been assigned to the cluster of *reports* when clustering is applied.

```
SVM classification

 a  b  c   <-- classified as
 34  9  6 |  a = tutorials
 12 27  9 |  b = reports
  6  8 32 |  c = research articles

Cluster assignments

 a   b  c    <-- assigned to cluster
 26  15  8  | a = tutorials
 10  31  7  | b = reports
  9   7 30  | c = research articles
```

Figure 7.4: Confusion matrices for SVM and K-means.

7.5 Summary of the experiments with the *articles* class

The most striking conclusion that can be drawn from the experiments in this chapter is that feature sets are highly dependent on target classes for classification. The feature ranking procedures applied show that several features of the base set $\mathtt{x}_{base}$ found effective in the previous chapter now turn out to be rather ineffective.

Given that the three broad classes attempted to identify in this chapter all have a probability of occurrence in the data set of slightly above 0.3, their performance figures are generally fairly promising. The genre space considered here is thus worthy of more investigations, even though the accuracy figures around 60$\%$ are far from impressive. One way of doing this is of course to revisit the annotation and reconsider the label assignments, possibly with some interhuman experiments, in order to determine the complexity or vagueness of this grouping of documents in relation to the notion of genre. The distinction between reports and research articles may not be as sharp with respect to intentional purpose when it comes to certain domains, as many research articles intends to present ongoing or finished research projects. For instance, the description of the "EuTEACH" project in `articleprojdesc24` is of an ongoing project, but is assigned not to the research article class, but to the report class. The purpose is highly similar for the *D-LIB Magazine* article that reports preliminary results of an ongoing research project to "teach incoming undergraduate students information literacy skills" (`articleAR2`). The difference is subtle and appears to be based on how much care is taken when preparing the text.

However, the attempt to assess the impact of the speech act feature set $\mathtt{x}_{act}$ and the document structure feature set $\mathtt{x}_{struct}$, that was the main issue for this chapter, remains inconclusive. The figures attained are summarized in Table 7.3.

The somewhat discouraging results, with respect to the additional feature sets, should probably not be overestimated before a more thorough investigation of the data set has been made. After all, the feature evaluation performed in this chapter showed that the impact of features is quite different from when other genre spaces are considered. For instance, reference lists and abstracts can intuitively be expected to be fruitful for discriminating scholarly material from non-scholarly material, if they can be algorithmically identified.

			F-scores		
No. of Features	Model	Accuracy	Tutorials	Reports	Research articles
$\lvert\mathbf{x}_{base}\rvert$	SVM	61.7 ± 0.9	64.7 ± 0.9	55.8 ± 1.9	65.1 ± 1.0
$\lvert\mathbf{x}_{base}\rvert$	7-NN	62.1 ± 0.9	54.7 ± 1.5	64.0 ± 1.2	66.8 ± 1.2
$\lvert\mathbf{x}_{base}\rvert$	K-means	55			
$\lvert\mathbf{x}_{base} \cup \mathbf{x}_{act}\rvert$	SVM	59.2 ± 1.0	61.0 ± 1.3	51.8 ± 1.3	61.5 ± 1.4
$\lvert\mathbf{x}_{base} \cup \mathbf{x}_{act}\rvert - 3$	SVM	59.0 ± 1.0	62.4 ± 1.4	52.1 ± 1.1	62.1 ± 1.6
$\lvert\mathbf{x}_{base} \cup \mathbf{x}_{act}\rvert - 3$	K-means	60			
$\lvert\mathbf{x}_{base} \cup \mathbf{x}_{struct}\rvert$	SVM	62.7 ± 1.3	64.8 ± 1.9	56.7 ± 1.7	65.9 ± 1.4
$\lvert\mathbf{x}_{base} \cup \mathbf{x}_{struct}\rvert - 6$	SVM	62.8 ± 1.1	65.2 ± 1.2	57.8 ± 1.4	65.9 ± 1.4
$\lvert\mathbf{x}_{base} \cup \mathbf{x}_{struct}\rvert - 6$	K-means	60			

Table 7.3: Summary of performance measures for the *articles* data set.

Part III

Conclusions and Discussions

Chapter 8

Concluding discussion

This work has been carried out from a starting point of the conception of genre as social action, and placed an interest in how this conception could be perused for the development of knowledge organisation theory and its application on classification tasks. A number of research questions were tentatively posed at the outset of this work, guided by these objectives. This chapter will summarize and discuss whether and how this investigation has contributed with anything substantial from which we can draw conclusions.

The first question posed was how genre is **conceptualized within LIS, linguistics and related disciplines, especially with respect to classification purposes?**

There is, of course, no short answer to this question. The question has been explored first in Chapter 2, and then partly in Chapter 4. The most striking feature of the varying understandings encountered is that there seems to be no clear interdisciplinary agreement on what the defining characteristics of a genre are. Even though most attempts to clarify the concept of genre in varying classification contexts often refer to the social understandings of Orlikowski and Yates, or of Swales and Miller, this is often superficial and the use of the concept is most commonly related to artefactual form, rather than to a document's context.

In *library classification* contexts, the notion of genre is largely undertheorized. When the word genre is used, it is taken for granted that its meaning is generally understood, but attempts to define genre are remarkably scarce and a common artefactual form is often stressed as the defining characteristic. However, recent years have shown an increasing interest in genre as such, especially regarding genres on the web.

In *linguistics*, the word genre is often avoided. However, in the subdisciplines of sociolinguistics, text linguistics, and neighbouring domains concerned not only with the levels of paragraphs, sentences, and words, the notions of register and text types turn out to be closely related to genre. Here we find an interest in the typified use of language in particular situational and communicational contexts, which is in essence the realization of genre as social action.

Text technology, and in particular *markup theory*, likewise avoid the word genre to a large extent. It is then in the notion of document types, often expressed in conjunction with SGML or XML applications, we find traces of a theory of genre. Document types are prescriptive and express conventions related to the compositional structure of a document, which can be seen as realizations of coarse-grained genres.

The most valuable empirical knowledge on genres cab probably be found in register studies or in studies of the systemic-functional school (see Section 2.2), even though the word register is to some extent almost as elusive as genre. But to equate register with linguistic patterns that bear a one-to-one relationship with genres is probably a mistake. A set of observable patterns in linguistic expressions, or in extra-linguistic expressions, is not the same as a genre even though it may be the outcome of a particular genre. However, this is the approximating assumption behind any approach to applied classification along genre dimensions — inferring genre from the artefactual patterns of its documents. The names given to classes of documents, formed on the basis of their appearance, should likewise not straightforwardly be taken as unambiguous names of genres. Therefore, genre classification may fail just because we have confused a class of similar artefacts with a genre. The three obviously different document structures found in the KI-04 class *discussion pages* seem to confirm this — at least with respect to some features (see Section 5.3.2). It can therefore be questioned whether it is strictly correct to talk about genre classification when the similarity measures of the harmonic quality introduced in Section 3.1 are defined by means of artefactual features only, and not by means of a contextual configuration. However, with such a strict view in mind, no algorithms would be possible.

The second question posed was **how different applications of document genre classification are realized, or how they can be effectively realized?** It is clear right from the start that algorithms rest on feature derivation from observable patterns in the artefacts themselves.

This question has been partly explored first in Chapter 3, then in Chapter 4, and finally in the experimental part of this thesis. It has been shown that there are many ways of realizing classification, and that the different ways are seldom comparable to one another. Different genre spaces and different feature sets have been used in previous research. This is a drawback, as all of the previous research efforts are then fairly inconclusive with respect to whether document genre classification is successful for real world applications or not. In addition, the definitions of genres are often rather vague. On the other hand, the inconclusiveness of previous research is also promising, because most experiments show that algorithms at least perform better than pure chance, which is also the case for the experiments of this work.

There are still several kinds of features that are left unexplored, such as speech act categories, which reflects a different kind of performative language use that can, at least in theory, be expected to indicate the particular illocutionary acts that characterise the purposes of language use. In addition, the kind of extra-linguistic structures conveyed by markup languages and techniques for visual formatting have been partially explored but deserve more attention. The experiments of this work have not shown any success for these kinds of features, when applied on a fine-grained genre space of articles. Despite this fact the special encodings and well-recognizable contents of lists of references, for instance, are not likely to be that difficult to identify over more coarse-grained genre spaces, and are probably fairly discriminative for scholarly material in general.

Another observation worth mentioning, with regards to features employed in previous research, is that in recent years we have seen many attempts to incorporate computationally expensive natural language processing tasks as a preprocessing requirement. This is a parallel to what has happened with the development of information retrieval research. But as it is hard to state in general terms that such things as lemmatization and parts-of-speech tagging is always worth the cost in topical information retrieval, the same may hold true for document genre classification. The results of this work reported in Section 6.2 indicate that for some simple tasks one may very well be content with more simple features, and that, in this case, words anticipated as discriminative, may be fed directly to a search engine.

Another distinguishing feature of most previous research, the experiments of this work, and the ways in which librarians apply classification according to genre, is that disjoint classification is fostered. There are some exceptions, but in general, a document is considered to adhere to just one genre (or class). This implies a deterministic thinking, where a document is part of only one situational and sociocultural context and performs just one function. This is, of course, a pragmatic simplification. The production of a document is probably determined by a single primary context, but its future use may be manifold, and thus its function. Overlapping document genre classification, which has been tried out, is a compelling task, but puts special requirements on implementing the classification models that are inherently inclined for disjoint classification.

Another disturbing fact about classification models is that they are constructed with an exhaustive classification in mind, which implies a predefined genre space, and instances which do not fit into any of these genres pose serious problems. This issue can sometimes be solved by assuming a tailing class which groups together all instances that do not fit in any of the other classes. However, such a class of unfit documents would result in a class where the feature-values may be widely dispersed in the feature space, and could thus confuse any algorithm.

The summarizing conclusion to be drawn from this second question is that the classification along genre dimensions can be performed as a *human classification* task, but that it then suffers from a vague definition of genre and the almost impossible task of naming genres. *Supervised classification* is an alternative solution, where classification always depends on the feature sets and algorithms employed, which generally need to be tailored with respect to the target classes of interest. Moreover, the number and nature of target genre classes are important for its success. The number of genres targeted usually needs to be low. In addition, a supervised classification rests on the assumption of user feedback and/or preclassified corpora, it suffers from the same shortcomings as does human classification. *Unsupervised classification* seems to be a good choice as long as the target genres are fairly distant from each other, meaning that, the artefacts of each class are distinct from the members of any other class. Otherwise it seems that the artefacts of genres are too unpredictable with respect to the feature sets employed in previous research.

The third, and more critical, question posed was to **what extent these approaches** [in document genre classification] **comply with an understanding of genre as social action?**.

In the approaches to classification based on algorithms, the simple answer is that they do not comply at all, as algorithms need observable patterns in the form of features to rely on. The target genres used in previous research are mostly defined without any clear reference to any situational and communicational context. This has been further elaborated in Section 4.1. A very similar argument goes for the approaches to human classification that have been recapitulated in this thesis, mainly in sections 2.1.3 and 2.1.2. Genres are not seldom intermingled with what would rather be termed media types, carrier types, or forms of expression.

Partly, it has already been stated above that there is a risk that the necessary and inherent focus on observable and extractable features is an essential and approximative simplification. However, even though most of the previous researchers declare a dependency on the genre theoretical views of Orlikowsky, Miller or Swales, there is almost always a tendency towards interpreting genres as equal to classes of documents where artefactual form is the dominant property. This tension between theory and application is in fact noted by James R. Martin in a preface to a just recently published antology on “genres on the web”:

> What if genres cannot be robustly characterised on the basis of just a few easily computable formal features? What if a flat approach to contextual variables and representational features simplifies research to the point where it is hard to see how the texts considered could have evolved as realisations of the genres members of our culture use to live. (Mehler et al., 2010, p. vii)

The final question posed was: **How do different definitions of genre spaces, classification models and document features influence document genre classification?** It has been necessary to take a pragmatic stance with respect to this question, as it is not possible to arrive at a completely satisfactory answer . The ways in which genres can be defined vary a lot. The number of genre spaces of is practically infinite. The amount of classification models to choose from is far too large, and previous research points in a direction towards the SVM model as being superior. As for document feature sets, the same as is said for the number of genre spaces holds for feature sets as well.

Therefore, the answer to this question is generalizable only insofar as is possible with respect to a) the empirical data chosen, b) the features actually derived, and c) a few configurations of the target genre space considered to be of interest.

Three models for classification have been applied. The experiments show that Support Vector Machines (SVM) almost always outperformed k-Nearest-Neighbor classification, even though no attempt to tune the parameters of SVM has been made. As for K-means clustering, where, strictly speaking, no learning actually takes place, the algorithm generally performs much worse than the other two, when its clusters are evaluated against the gold standard. The only exception is for the relatively simple task of discriminating between *articles* and *discussion pages*. This comes as no surprise and only corroborates what other research has shown.

The number of target classes for a classification task generally affects the performance, so that an increasing number of classes decreases the performance. This is no surprise either, as the number of classes is bound to increase the mathematical measure of entropy if the classes are fairly evenly distributed. However, the nature of the target classes is equally important, so that the performance cannot be considered a simple function of entropy. The performance differences between dividing the full KI-04 set into 8 classes and the *articles* class into 3 more fine-grained subclasses are small, even though it should be greater if entropy is considered. The differences when comparing the tasks of weeding out *articles* from *help pages* and *articles* from *discussion pages* are striking, and are most probably dependent on the nature of the classes of documents of each genre in relation to the feature set derived. Thus, entropy cannot be considered the only matter that affects performance. It also indicates that some features seem to be more successful for some tasks.

The broad group of genres connected to the *articles* class is what has been most fully explored in this work. It is clear from the first set of experiments that it is easier to gain a high precision in weeding out documents of the *articles* class from the rest of the *KI-04* classes, compared to recognizing them all, and to thus gain a good recall. This may be due to several reasons. First, one may suspect that this class is far too coarse-grained (and thus heterogeneous), or that it suffers from many misclassifications in the annotation. A slight increase in performance may be discerned when the class is reexamined for the removal of misclassifications, but more striking differences occur when the class is extended by documents more carefully chosen to fit the "center" of the class, being somewhat prototypical. A more accurate conclusion is therefore to state that it is not the low granularity or a few misclassifications that is the main problem, but an insufficient amount of training data that support the algorithm *in combination* with a low granularity and noise.

Turning to the more fine-grained experiments with the *articles* class, they confirm that the amount of training data is a crucial issue. Even though a classification of the instances of *articles* class into three classes, broadly labelled *tutorials*, *reports* and *research articles*, performs better than pure chance, the results gained for each class are similar to that of the eight-class task, but have confidence intervals that are fairly high. The differences between supervised and unsupervised approaches are also much smaller than for the experiments with larger data sets, which may suggest that the features themselves, taken as a whole, are fairly accurate for the task. Thus, this must be considered a more difficult classification task, compared to when the granularity is low. One reason may very well be that the expected structure of a technical report (as long as a standard structure is obeyed) is fairly similar to the expected structure of a research article in a technical journal, and that the speech acts associated with the two genre typifications are equally similar. The misclassifications are then due to similarities inherent in the genres. It is likely that we need to know more about which artefactual features distinguish these genres from each other, in order to extend the feature set.

The impact of document structure features is a final issue to consider. Initially, it was assumed that the difference between, for instance, a *research article* and a *tutorial* is in terms of purpose, the stance of the author towards his or her audience, the "tenor" (if applying the concept of the systemic-functional school) , the illocutionary power put into language use, and a more rigid document structure in the case of *research articles*. Thereby, the addition of features that, so to say, model speech acts and document structures, although on a superficial level, should significantly improve the performance. This cannot at all be concluded. However, when previous research is considered, together with some of the experiments of this work, it is clear that the impact of different kinds of features is highly dependent on the character of the target classes. The most striking example is for the experiments in Section 6.2, where a few tailored features, derived from the contents of the titles only, radically increase the performance. One may therefore say that these features should be further investigated before dismissing them as being inadequate. In addition, the rather conservative approach in choosing the lexical repertoire for each speech act feature could be reconsidered, as to including a repertoire that mirrors contemporary language use.

In order to boil down the answers to the last question as much as possible, the following conclusions can be drawn:

- When the *number of target genres* is increased, the classification task becomes computationally more cumbersome. This is corroborated in the experiments of this work, but is also shown to be dependent on the nature of the target genres.

- When the *distance between target genres*, in terms of document instances prototypical of a genre, is sufficiently large, the task of classification is more robust, shown most clearly by the experiments reported in Section 6.3, and in particular on page 197. In these cases, *unsupervised classification* seems plausible.

- When the *number of features* is increased, the classification performance increase as well. However, the effect of individual features seems to be highly dependent on the target genre space, judging from the experiments of this work as a whole, as well as from previous research.

- *Hand-tailored features*, based on assumption of word occurrences, are highly plausible as long as the target genres allow for that kind of assumption.

- The extension of the source corpus with somewhat *prototypical instances* is beneficial, as shown in Table 6.5.

- *Speech act features* have not been proven effective in these experiments. In fact, in the three-class problem of the subdivision of *articles*, they seem to harm the performance.

- *Document structure features* seem to show no effect at all on the same three-class problem.

Chapter 9

Suggestions for further research

The experimental outcome of this work has not provided any exhaustive answers to the questions about whether classification along genre dimensions has the potentials necessary to perform successfully in real world applications, and about whether the two fairly novel feature sets really are effective in classification tasks. Theoretically, though, both classes of verb forms and certain document structure element types remain promising and the empirical inconclusiveness may be explained by a few disadvantageous factors of the experiments. The size of the data is probably the most obvious one. The two feature sets need further investigations where, for instance, verb forms are derived from their cotext by means of e.g. *n*-grams. Moreover, a repertoire of verb forms justified by means of an analysis of modern functional language use could be useful. The catalog of speech-act verbs may very well be outdated.

This points to the most obvious need for future research — **corpora** that reflect contemporary language use, compiled and annotated for research in particular genre variation, and more comprehensive than the existing ones. Such corpora would most preferrably show the following characteristics:

- Being compiled for genre spaces that are crucial for particular information seeking tasks, and therefore preferrably developed in cooperation between linguists, library and information scientists, and computer scientists

- Being annotated with respect to different justified granularities of genre spaces, as well as documenting the genres in terms of elaborated definitions and descriptions of their situational and cultural contexts
- Being annotated with respect to different linguistic characteristics, such as parts-of-speech, as well as with respect to paralinguistic characteristics, such as document structural elements
- Supporting stand-alone annotation, preferrably in XML-compliant form, that would allow parallel annotation of different kinds

Future research could then be guided by the following objectives:

- Further investigating how to infer document structure features and put them to use in classification tasks
- Exploring the relationships between topics, domains, and genres
- Finding novel features of different kinds that are justified with respect to a view on language use as reflecting speech acts, maybe by means of factor analysis or similar techniques
- Evaluating real world applications based on user feedback
- Further investigating the intersubjectivity of classification along genre dimensions, in terms of human understandings of the nature of genres
- Developing methods for extended corpus annotation by means of user feedback

A prerequisite for such research to be successful is probably that it is interdisciplinary designed with contributions from e.g. sociolinguists, computational linguists, and library and information scientists.

Bibliography

Aha, D. W., Kibler, D., & Albert, M. (1991). Instance-based learning algorithms. *Machine Learning*, 6, 37–66.

Alpaydin, E. (2004). *Introduction to Machine Learning*. Cambridge, Mass.: MIT Press.

American National Standards Institute, Ed. (2005). *Scientific and Technical Reports – Preparation, Presentation, and Preservation. ANSI/NISO Z39.18-2005*. Bethesda, Maryland: NISO Press.

Andersen, J. (2004). *Analyzing the role of knowledge organization in scholarly communication: An inquiry into the intellectual foundation of knowledge organization*. PhD thesis, Department of Information Studies, Royal School of Library and Information Science. Copenhagen University.

Argamon, S., Koppel, M., & Avneri, G. (1998). Routing documents according to style. In *Proceedings of the First International Workshop on Innovative Internet Information Systems*.

Austin, J. L. (1975). *How to do things with words: the William James Lectures delivered at Harvard University in 1955*. Oxford: Oxford University Press, 2 edition.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press.

Bates, M. J. (2002). Speculations on browsing, directed searching, and linking in relation to the bradford distribution. In H. Bruce, R.

Fidel, P. Ingwersen, & P. Vakkari (Eds.), *Emerging Frameworks and Methods: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS 4)* (pp. 137–150).: Libraries Unlimited.

Batley, S. (2005). *Classification in Theory and Practice*. Oxford: Chandos.

Bazerman, C. (1989). *The informed writer: using sources in the discipline*. Boston, Mass: Haughton Mifflin Co, 3 edition.

Bazerman, C. (1994). Systems of genres and the enactment of social intentions. In A. Freedman & P. Medway (Eds.), *Genre and the New Rhetoric*, Critical Perspectives on Literacy and Education (pp. 79–101). London: Taylor & Francis.

Beghtol, C. (1998). General classification systems: structural principles for multidisciplinary specification. In S. P. W. Mustafa el Hadi, J. Maniez (Ed.), *Structures and relations in knowledge organization: Proceedings of the 5th International ISKO Conference, Lille, 25-29 August 1998* (pp. 89–96). Würzburg: Ergon.

Beghtol, C. (2001). The concept of genre and its characteristics. *Bulletin of the American Society for Information Science and Technology*, 27(2), 1–5.

Berners-Lee, T. (1989). Information management: A proposal. The original proposal of the WWW.

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

Biber, D. (1994). An analytical framework for register studies. In D. Biber & E. Finegan (Eds.), *Sociolinguistic Perspectives on Register* (pp. 31–56). Oxford University Press.

Biber, D. & Ferguson, C. A., Eds. (1994). *Sociolinguistic Perspectives on Register*, chapter Introduction: Situating Register in Sociolinguistics, (pp. 3–12). Oxford University Press.

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

Bodoff, D. (2006). Relevance for browsing, relevance for searching. *Journal of the American Society for Information Science*, 57(1), 69–86.

Boese, E. & Howe, A. H. (2005). Effects of web document evolution on genre classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (pp. 632–639).: ACM.

Boese, E. S. (2005). Stereotyping the web: Genre classification of web documents. Master's thesis, Department of Computer Science. Colorado State University, Fort Collins, Colorado.

Bogdanov, A. D. & Worring, M. (2001). Fine-grained document genre classification using first order random graphs. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*.

Briet, S. (1951). *Qu'est-ce que la documentation?* Paris: EDIT.

Broughton, V. (2004). *Essential Classification*. London: Facet Publishing.

Brown, J. S. & Duguid, P. (1996). The social life of documents. *First Monday*, 1(1).

Buchanan, B. (1979). *Theory of Library Classification*. Outlines of Modern Librarianship. London: Clive Bingley.

Buckland, M. (1991). What is a 'document'? *Journal of the American Society of Information Science*, 48(9), 804–809.

Buckland, M. K. (2007). Naming in the library: Marks, meaning and machines. In C. Todenhagen & W. Thiele (Eds.), *Nominalization, Nomination and Naming in Texts*. Tübingen: Stauffenberg.

Chan, L. M. (1990). *Library of Congress Subject Headings - Principles of Structure and Policies for Application*. Library of Congress.

Chiang, J., Hao, P., & Tu, Y. (2007). Hierarchically SVM classification based on support vector clustering method and its application to document categorization. *Expert systems with applications*, 33, 627–635.

Comaromi, J. P., Beall, J., Matthews, W. E., & New, G. R., Eds. (1989). *Dewey Decimal Classification and Relative Index*. Forest Press, 20 edition.

Coombs, J. H., Renear, A. H., & DeRose, S. J. (1987). Markup systems and the future of scholarly text processing. *Communications of the ACM*, 30(11), 933–947.

Crowston, K. (2010). Internet genres. In M. J. Bates & M. N. Maack (Eds.), *Encyclopedia of Library and Information Sciences* (pp. 2983–2995). Taylor and Francis, 3 edition.

Crowston, K. & Kwasnik, B. H. (2003). Can document-genre metadata improve information access to large digital collections. *Library Trends*, 52(2), 345–361.

Crowston, K. & Kwasnik, B. H. (2004). A framework for creating a facetted classification for genres: Addressing issues of multidimensionality. In *Proceedings of the 37th Hawaii International Conference on System Sciences*: IEEE.

Crowston, K. & Williams, M. (2000). Reproduced and emergent genres of communication on the world wide web. *Information Society*, 16(3), 201–216.

Cutter, C. A. (1904). *Rules for a Dictionary Catalogue*. USGPO, 4 edition.

Daelemans, W., Zavrel, J., van der Sloot, K., & von den Bosch, A. (2004). *TiMBL: Tilburg Memory-Based Learner. Reference Guide*.

Technical report ILK 04-02, Tilburg University. Induction of Linguistic Knowledge.

Dahlberg, I. (1978). *Ontical Structures and Universal Classification*. Bangalore: Sarada Ranganathan Endowment for Library Science.

Dahlström, M. (2002a). Nya medier, gamla verktyg. *Human IT*, (4), 71–116.

Dahlström, M. (2002b). When is a webtext? *TEXT Technology*, 11(1), 139–161.

Dahlström, M. (2006). *Under utgivning*. PhD thesis, Inst. f. Biblioteks- och Informationsvetenskap. Göteborgs Universitet.

Dahlström, M. & Gunnarsson, M. (1999). DA draws a circle : on document architecture and its relation to library and information science education and research. *Information Research*, 5(2).

Dewdney, N., VanEss-Dykema, C., & McMillan, R. (2001). The form is the substance: Classification of genres in text. In *ACL Workshop on Human Language Technology and Knowledge Management* (pp. 142–149).

Drake, M. A., Ed. (2003). *Encyclopedia of Library and Information Science*. New York: Dekker, 2 edition.

Dubin, D., Renear, A., Sperberg-McQueen, C. M., & Huitfeldt, C. (2003). A logic programming environment for document semantics and inference. *Literary & Linguistic Computing*, 18(1), 39–47.

Dura, E., Olsson, B., Erlendsson, B., & Gawronska, B. (2006). Information fusion in pathway evaluation: Encoding of relations in biomedical texts. In *9th International Conference on Information Fusion, Florence, 10-13 July 2006* (pp. 240–247).

Ellis, D. (1996). *Progress & problems in information retrieval*. New Horizons in Information Retrieval. Library Association Publishing, 2 edition.

Elsas, J. & Efron, M. (2004). HTML tag based metrics for use in web page type classification.

Englund, B. & Ledin, P., Eds. (2003). *Teoretiska perspektiv på sakprosa*. Lund: Studentlitteratur.

Feather, J. & Sturges, P., Eds. (1997). *International Encyclopedia of Information and Library Science*. London: Routledge.

Ferguson, C. A. (1994). Dialect, register, and genre: Working assumptions about conventionalization. In D. Biber & E. Finegan (Eds.), *Sociolinguistic Perspectives on Register* (pp. 15–30). Oxford University Press.

Finn, A. & Kushmerick, N. (2003). Learning to classify documents according to genre. In *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*.

Folch, H., Heiden, S., Habert, B., Illouz, G., Fleury, S., Lafon, P., Nioche, J., & Prévost, S. (2000). TyPTex: Inductive typological text classification by multivariate statistical analysis for NLP systems tuning/evaluation. In *LREC 2000*.

Foskett, A. C. (1996). *The Subject Approach to Information*. London: Library Association Publ, 5 edition.

Francke, H. (2005). What's in a name? contextualizing the document concept. *Literary & Linguistic Computing*, 20(1), 61–69.

Francke, H. (2008). *(Re)creations of Scholarly Journals: Document and Information Architecture in Open Access Journals*. PhD thesis, Inst. f. Biblioteks- och Informationsvetenskap. Göteborgs Universitet.

Freedman, A. & Medway, P., Eds. (1994). *Genre and the New Rhetoric*. London: Taylor & Francis.

Frohmann, B. (1992). The power of images: a discourse analysis of the cognitive viewpoint. *Journal of Documentation*, 48(4), 365–386.

Frohmann, B. (1994). The social construction of knowledge organisation: The case of Melvyl Dewey. In *Knowledge Organization and Quality Management. Proceedings of the Third International ISKO Conference, 20-24 June 1994*. (pp. 109–117).: Indeks verlag.

Frohmann, B. (2004). Documentation redux: Prolegomenon to (another) philosophy of information. *Library Trends*, 52(3), 387–408.

Goldstein, J. & Evans Sabin, R. (2006). Using speech acts to categorize email and identify email genres. In *Proceedings of the 39th Hawaii International Conference on System Sciences*: IEEE.

Grepstad, O. (1997). *Det litterære skattkammer: sakprosans teori og retorikk*. Oslo: Det Norske Samlaget.

Görlach, M. (2004). *Text Types and the History of English*. Trends in Linguistics. Studies and Monographs; 139. Berlin: Mouton de Gruyter.

Haas, S. & Grams, E. (2000). Readers, authors, and page structure: a discussion of four questions arising from a content analysis of web pages. *Journal of the American Society for Information Science*, 51(2), 181–192.

Halliday, M. A. K. & Hasan, R. (1989). *Language, Context and Text; Aspects of Language in a Social-semiotic Perspective*. Oxford: Oxford University Press, 2 edition.

Halliday, M. A. K. & Martin, J. R. (1993). *Writing science : literacy and discursive power*. Critical perspectives on literacy and education. London: Falmer.

Halliday, M. A. K. & Matthiessen, C. M. I. M. (2004). *An Introduction to Functional Grammar*. London: Arnold, 3 edition.

Hansson, J. (1999). *Klassifikation, bibliotek och samhälle*. Phd, Inst. f. Biblioteks- och Informationsvetenskap. Göteborgs Universitet.

Harter, S. (1986). *Online information retrieval: concepts, principles, and techniques*. Library and Information Science Series. Academic press.

Hellspong, L. & Ledin, P. (1997). *Vägar genom texten : handbok i brukstextanalys*. Lund: Studentlitteratur.

Hjørland, B. (1997). *Information Seeking and Subject Representation*. New Directions in Information Management, No. 34. New York: Greenwood Press.

Hjørland, B. (1998). Information retrieval, text composition, and semantics. *Knowledge Organization*, 25(1/2).

Hjørland, B. (2002). Domain analysis in information science: Eleven approaches - traditional as well as innovative. *Journal of Documentation*, 58(4), 422–462.

Hjørland, B. (2006). Lifeboat for knowledge organisation.

Hjørland, B. & Nicolaisen, J., Eds. (2005-2006). *The Epistemological Lifeboat Epistemology and Philosophy of Science for Information Scientists*. Danmarks Biblioteksskole.

Honkaranta, A. (2003). *From genres to content analysis: experiences from four case organizations*. PhD thesis, Department of Computer Science and Information Systems. University of Jyväskylä. Jyväskylä Studies in Computing ; 31.

Hu, J., Kashi, R., & Wilfong, G. T. (1999). Document classification using layout analysis. In *DEXA Workshop* (pp. 556–560).

Hunter, E. J. (2002). *Classification made simple*. Aldershot: Ashgate.

Hutchins, W. J. (1978). The concept of 'aboutness' in subject indexing. *ASLIB Proceedings*, 30(5), 172–181.

Hymes, D., Ed. (1974). *Foundations in Sociolinguistics*. Philadelphia, PA: Univ. of Pennssylvania Press.

Ihlström, C. & Åkesson, M. (2004). Genre characteristics: a front page analysis of 85 swedish online newspapers. In *Proceedings of 37' Hawaii International Conference on Systems Science*: IEEE Press.

Illich, I. (1993). *In the Vineyard of the Text; A Commentary to Hugh's Didascalion*. Chicago: The University of Chicago Press.

Jacob, E. K. (2004). Classification and categorization. *Library Trends*, 52(3), 515–540.

Jarneving, B. (2005). *The Combined Application of Bibliographic Coupling and the Complete Link Cluster Method in Bibliometric Science Mapping*. Phd thesis, Inst. f. Biblioteks- och Informationsvetenskap. Göteborgs Universitet.

Jurafsky, D. & Martin, J. H. (2000). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence. Upper Saddle River: Prentice Hall.

Karlgren, J. (2000). *Stylistic experiments for information retrieval*. PhD thesis, Department of Linguistics. Stockholm Univ.

Kaufer, D., Geisler, C., Ishizaki, S., & Vlachos, P. (2005). Textual genre analysis and identification. In *Ambient Intelligence for Scientific Discovery* (pp. 129–151).: Springer.

Kent, A., Ed. (1968-2003). *Encyclopedia of Library and Information Science*. New York: Dekker.

Kessler, B., Nunberg, G., & Schütze, H. (1997). Automatic detection of text genre. In *Proceedings of the 35th annual meeting of the Association for Computational Linguistics and the 8th meeting of the European Chapter of the Association for Computational Linguistics* (pp. 32–38). San Francisco: Morgan Kaufmann Publishers.

Kim, Y. & Ross, S. (2007). 'the naming of cats': Automated genre classification. *The International Journal of Digital Curation*, 2(1), 49–61.

Kim, Y. & Ross, S. (2008). Examining variations of prominent features in genre classification. In *Proceedings of the 41st Hawaii International Conference on System Sciences*: IEEE.

Kim, Y.-J. & Biber, D. (1994). A corpus-based analysis of register variation in korean. In D. Biber & E. Finegan (Eds.), *Sociolinguistic Perspectives on Register* (pp. 157–181). Oxford University Press.

Koch, T., Ardö, A., & Golub, K. (2004). Log analysis of user behaviour in the renardus web service. In *Human Information Behaviour & Competences For Digital Libraries. Proceedings, Libraries in the Digital Age (LIDA).* (pp. 175–177).

Koch, T., Day, M., Brümmer, A., Hion, D., Peereboom, M., Poulter, A., & Worsfold, E. (1997). *Specification for Resource Description Methods: Part 3: The Role of Classification Schemes in Internet Resource Description and Discovery*. DESIRE.

Kwasnik, B. H., Crowston, K., Nilan, M., & Roussinov, D. (2001). Identifying document genre to improve web search effectiveness. *Bulletin of the American Society for Information Science and Technology*, 27(2).

Lancaster, F. W. (1998). *Indexing and Abstracting in Theory and Practice*. London: Library Association Publishing, 2 edition.

Langridge, D. W. (1989). *Subject Analysis: Principles and Procedures*. London: Bowker-Saur.

Latour, B. (1986). Visualization and cognition: Thinking with eyes and hands. *Knowledge and Society*, 6, 1–40.

Lavesson, N. (2006). *Evaluation and Analysis of Supervised Learning Algorithms and Classifiers*. Licentiate thesis, Blekinge

Institute of Technology. Department of Systems and Software Engineering.

Ledin, P. (1999). *Texter och textslag – en teoretisk diskussion*. Rapporter från projektet Svensk sakprosa; 27. Lund: Institutionen för nordiska språk.

Lee, D. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & and Technology*, 5(3), 37–72.

Lee, Y.-B. & Myaeng, S. H. (2002). Text genre classification with genre-revealing and subject-revealing features. In *SIGIR '02* (pp. 145–149).

Leininger, K. (2000). Indexer consistency in PsycINFO. *Journal of Librarianship and Information Sceince*, 32(4), 4–8.

Levy, D. M. (2000). Where's waldo? : Reflections on copies and authenticity in a digital environment. In *Authenticity in a Digital Environment, Washington, D.C.* (pp. 24–31).: Council on Library & Information Resources.

Levy, D. M. (2003a). Documents and libraries: A sociotechnical perspective. In A. Peterson Bishop, N. A. Van House, & B. P. Buttenfield (Eds.), *Digital Library Use*. Cambridge, Mass.: MIT Press.

Levy, D. M. (2003b). From documents to information: A historical perspective. In *The Document Academy Meeting, Berkeley, Ca*.

Library of Congress (1999 (updated 2004)). MARC 21 format for bibliographic data: Field list.

Lim, C. S., Lee, K. J., & Kim, G. C. (2005). Multiple sets of features for automatic genre classification of web documents. *Information Processing and Management 41 (2005) 1263–1276*, 41, 1263–1276.

Lubell, J. (2001). Architectures in an XML world. *Markup Languages*, 3(4).

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge Press.

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Mass: MIT Press.

Marchionini, G. (1995). *Information Seeking in Electronic Environments*. Cambridge Series on Human-Computer Interaction. Cambridge: Cambridge University Press.

Mayes, P. (2003). *Language, social structure and culture*. Amsterdam: John Benjamins.

McEnery, T. & Wilson, A. (2001). *Corpus Linguistics*. Edinburgh textbooks in empirical linguistics. Edinburgh: Edinburgh Univ. Press, 2 edition.

McKelvie, D., Brew, C., & Thompson, H. S. (1998). Using SGML as a basis for data-intensive natural language processing. *Computers and the Humanities*, 31, 367–388.

Megyesi, B. (2002). *Data-Driven Syntactic Analysis*. PhD thesis, Dep. of Speech, Music and Hearing. KTH, Stockholm.

Mehler, A. (2007). Structure formation in the web. toward a graph-theoretical model of hypertext types. In A. Witt & D. Metzing (Eds.), *Linguistic Modelling of Information and Markup Languages* (pp. 1–36). Springer.

Mehler, A., Geibel, P., & Pustylnikov, O. (2007). Structural classifiers of text types: Towards a novel model of text representation. *LDV Forum*, 22(2), 51–66.

Mehler, A., Sharoff, S., & Santini, M., Eds. (2010). *Genres on the Web: Computational Models and Empirical Studies*. Dordrecht: Springer.

Meyer zu Eissen, S. (2007). *On Information Need and Categorizing Search*. PhD thesis, Dep. of Computer Science. University of Paderborn, Germany, Paderborn.

Meyer zu Eissen, S. & Stein, B. (2004). Genre classification of web pages: User study and feasibility analysis. In P. G. Biundo S., Fruhwirth T. (Ed.), *Advances in Artificial Intelligence* (pp. 256–269). Berlin: Springer.

Miksa, F. L. (1992). The concept of the universe of knowledge and the purpose of LIS classification. In *Classification Research for Knowledge Representation and Organization: Proceedings of the 5th International Study Conference on Classification Research* (pp. 101–126).: Elsevier.

Miller, C. R. (1994). Genre as social action. In A. Freedman & P. Medway (Eds.), *Genre and the New Rhetoric*, Critical Perspectives on Literacy and Education (pp. 23–42). London: Taylor & Francis.

Mills, J. & Broughton, V., Eds. (1977). *Bliss Bibliographic Classification. Vol. 1: Introduction and Auxiliary Schedules*, chapter The Structure of a Bibliographic Classification, (pp. 35–47). Butterworths.

Mirkin, B. (2005). *Clustering for Data Mining: A Data Recovery Approach*. Computer Science and Data Analysis series ; 3. Boca Raton, FL: Taylor & Francis.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

Montesi, M. & Navarrete, T. (2008). Classifying web genres in context: A case study documenting the web genres used by a software engineer. *Information Processing and Management*, 44(4), 1410–1430.

Nicholas, D., Huntington, P., & Jamali Hamid R., Tenopir, C. (2006). Finding information in (very large) digital libraries: A deep log approach to determining differences in use according to method of access. *Journal of Academic Librarianship*, 32(2), 119–126.

Nunberg, G. (1999). *The Linguistics of Punctuation*. CSLI Lecture Notes ; 18. Stanford, CA: Center for the Study of Language and Information.

OCLC (2003). DDC 22 summaries.

Orlikowski, W. & Yates, J. (1994). Genre repertoire: the structuring of communicative practices in organizations. *Administrative science quarterly*, 39, 541–574.

Peels, A. J. H. M., Janssen, N. J. M., & Nawijn, W. (1985). Document architecture and text formatting. *ACM Transactions on Information Systems*, 3(4), 347–369.

Peterson Bishop, A., Van House, N. A., & Buttenfield, B. P., Eds. (2003). *Digital library use: social practice in design and evaluation*. Digital Libraries and Electronic Publishing. Cambridge, Mass.: MIT Press.

Power, R. & Scott, D. (1999). Using layout for the generation understanding or retrieval of documents : Papers from the 1999 AAAI fall symposium.

Power, R., Scott, D., & Bouayad-Agha, N. (2003). Document structure. *Computational Linguistics*, 29(2), 211–260.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Ranganathan, S. R. (1960). *Colon Classification*. Bombay: Asia Publ., 6 edition.

Ranganathan, S. R. (1967). *Prolegomena to Library Classification*. Bombay: Asia Publishing House, 3 edition.

Ranganathan, S. R. (1991). *Elements of Library Classification*. Bangalore: Sarada Ranganathan Endowment for Library Science, 2 edition.

Ranganathan, S. R. (1994). *Philosophy of Library Classification*. Bangalore: Sarada Ranganathan Endowment for Library Science.

Rauber, A. & Müller-Kögler, A. (2001). Integrating automatic genre analysis into digital libraries. In *JCDL '01, Roanoke, Virginia, USA*: ACM.

Raymond, D. K. (1992). Evolutions in typesetting systems. In *Ideas in Action:CASCON '92* (pp. 19–28). IBM Centre for Advanced Studies, IBM Canada Laboratory.

Rehm, G. (2002). Towards automatic web genre identification. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 4*: IEEE Computer Society.

Reitz, J. M. (2004). *Online Dictionary for Library and Information Science*. Westport, CT: Libraries Unlimited.

Renear, A., Dubin, D., Sperberg-McQueen, C. M., & Huitfeldt, C. (2002). Towards a semantics for XML markup. In R. Furuta, J. L. Maletic, & E. Munson (Eds.), *Proceedings of the 2002 ACM Symposium on Document Engineering, McLean, VA*. (pp. 119–126).: Association for Computing Machinery.

Renear, A. H. (2001). The descriptive/procedural distinction is flawed. *Markup Languages : Theory & Practice*, 2(4), 411–420.

Rosso, M. A. (2005). *Using Genre to Improve Web Search*. PhD thesis, School of Library and Information Science. University of North Carolina, Chapel Hill.

Samuelsson, J. (2008). *På väg från ingenstans: Kritik och emancipation av kunskapsorganisation för feministisk forskning*. Phd, Umeå Universitet.

Santini, M. (2004a). *Identifying Genres on the Web*. Technical Report ITRI-03-06, ITRI, Univ. of Brighton.

Santini, M. (2004b). *State-of-the-Art on automatic genre identification*. Technical Report ITRI-04-03, ITRI, Univ. of Brighton.

Santini, M. (2005). Genres in formation? an exploratory study of web pages using cluster analysis. In *Proceedings of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*.

Santini, M. (2006). Common criteria for genre classification: Annotation and granularity. In *Workshop on Text-Based Information Retrieval (TIR-06)* Riva del Garda, Italy.

Santini, M. (2007). *Automatic Identification of Genre in Web Pages*. Phd thesis, University of Brighton.

Santini, M., Sharoff, S., Rehm, G. & Mehler, A., Eds. (2008 –). WebGenreWiki : the wiki dedicated to Automatic Web Genre Identification.

Sasaki, F. & Pönninghaus, J. (2003). Testing structural properties in textual data: Beyond grammars. *Literary & Linguistic Computing*, 18(1), 89–100.

Schneider, J. W. (2004). *Verification of Bibliometric Method's Applicability for Thesaurus Construction*. PhD thesis, Royal School of Library and Information Science, Copenhagen.

Schryer, C. F. (2010). Genre theory and research. In M. J. Bates & M. N. Maack (Eds.), *Encyclopedia of Library and Information Sciences* (pp. 1931–1942). Taylor and Francis, 3 edition.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47.

Sebastiani, F. (2005). Text categorization. In A. Zanasi (Ed.), *Text Mining and its Applications to Intelligence, CRM and Knowledge Management* (pp. 109–129). WIT Press.

Sekine, S. (1998). *Corpus-based Parsing and Sublanguage Studies*. Ph. d., New York University. Computer Science Dep.

Shepherd, M., Watters, C., & Kennedy, A. (2004). Cybergenre: Automatic identification of home pages on the web. *Journal of Web Engineering*, 3, 236–251.

Sperberg-McQueen, C. M. (1991). Text in the electronic age: Textual study and text encoding with examples from medieval texts. *Literary and Lingustic Computing*, 6(1), 34–46.

Sperberg-McQueen, C. M. (2004). Classification and its structures. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A Companion to Digital Humanities* chapter 14. Oxford: Blackwell.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Text genre detection using common word frequencies. In *Proceedings of the 18th conference on Computational linguistics - Volume 2* (pp. 808–814). Luxembourg.

Stubbe, A. & Ringlstetter, C. (2007). Recognizing genres. In G. Rehm & M. Santini (Eds.), *Towards a Reference Corpus of Web Genres: Colloquium held in conjunction with Corpus Linguistics 2007, Birmingham, UK - July 27, 2007* (pp. 21–28).

Stubbe, A., Ringlstetter, C., & Goebel, R. (2007). Elements of a learning interface for genre qualified search. In G. Rehm & M. Santini (Eds.), *Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing. Proceedings* (pp. 21–28).

Study Group on the Functional Requirements for Bibliographic Records (1998). *Functional Requirements for Bibliographic Records*. München: K. G. Saur.

Subrahmanyam, B. (2006). Library of congress classification numbers: Issues of consistency and their implications for union catalogs. *Library Resources & Technical Services*, 50(2), 110–119.

Sukiasyan, E. (1998). Classification systems in their historical development: Problems of typology and terminology. In

Structures and Relations in Knowledge Organization. Proceedings of the Fifth International ISKO Conference, 25-29 August 1998. (pp. 72–79).: Ergon verlag.

Svenonius, E. (2000). *The Intellectual Foundation of Information Organization.* Digital Libraries and Electronic Publishing. MIT Press.

Swales, J. M. (1990). *Genre analysis; English in academic and research settings.* Cambridge: Cambridge Univ. Press.

Swanson, W. (2003). *Modes of Co-reference as an Indicator of Genre.* Linguistic Insights: Studies in Language and Communication; Vol. 12. Bern: Peter Lang.

Taylor, A. G. (1999). *The Organization of Information.* Englewood, Colorado: Libraries Unlimited.

Taylor, A. G. (2004). *The Organization of Information.* Westport, Connecticut: Libraries Unlimited, 2 edition.

Toms, E. G. (2001). Recognizing digital genre. *Bulletin of the American Society for Information Science and Technology*, 27(2).

Toms, E. G., Campbell, D. G., & Blades, R. (1999). Does genre define the shape of information: The role of form and function in user interaction with digital documents? In *ASIS '99: Proceedings of the 62nd ASIS Meeting, Washington, DC, October 31st-November 4th, 1999.* (pp. 693–704). Medford, NJ: Information Today.

van Rijsbergen, C. J. (1979). *Information Retrieval.* London: Butterworths, 2 edition.

Vaughan, M. W. & Dillon, A. (2006). Why structure and genre matter for users of digital information: A longitudinal experiment with readers of a web-based newspaper. *International Journal of Human-Computer Studies*, 64, 502–526.

Wastholm, P., Kusma, A., & Megyesi, B. (2005). Using linguistic data for genre classification. In *Advances in Artificial Language in Sweden. The Annual Swedish Artificial Intelligence and Learning Systems Event, SAIS-SSLS, April 2005* (pp. 173–176).

Werlich, E. (1976). *A Text Grammar of English*. Heidelberg: Quelle & Meyer.

Wilson, P. & Robinson, N. (1990). Form subdivisions and genre. *Library Resources & Technical Services*, 34(1), 36–43.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. London: Elsevier Science & Technology.

Wolters, M. & Kirsten, M. (1999). Exploring the use of linguistic features in domain and genre classification. In *Proceedings of the Ninth Conference on European chapter of the Association for Computational Linguistics* (pp. 142–149).: Association for Computational Linguistics.

Yule, G. (1996). *The study of language*. Cambridge: Cambridge University Press, 2 edition.

Ørom, A. (1997). Genre indenfor fagmedier. *Biblioteksarbejde*, (51/52), 7–29.

Ørom, A. (2003). Knowledge organization in the domain of art studies. *Knowledge Organization*, 30(3/4), 128–143.

Appendix A

The Definition of the Annotation Scheme

```
<!-- Elements common for both class definitions
     and annotations -->
<!ELEMENT collection (classdefs,instance*)>
<!ELEMENT note (#PCDATA)>

<!-- Class definitions. Note that the recursive
     definition of the class element allows for
     representing the classification scheme as
     a tree structure, though this is not being
     done here -->
<!ELEMENT classdefs (class*)>
<!ELEMENT class (name,description?,class*)>
<!ATTLIST class id ID #REQUIRED>
<!ELEMENT name (#PCDATA)>
<!ELEMENT description (definition?,function*,note?)>
<!ELEMENT function (#PCDATA)>
<!ELEMENT definition (#PCDATA)>

<!-- Annotations -->
<!ELEMENT instance (title?,community?,topic*,note?)>
<!ATTLIST instance id ID #REQUIRED
          class IDREF #IMPLIED>
<!ELEMENT community (#PCDATA)>
<!ELEMENT topic (#PCDATA)>
<!ELEMENT title (#PCDATA)>
```

Figure A.1: Document Type Definition for the annotation

```
/.../
<!DOCTYPE collection SYSTEM "coll.dtd" [
<!ENTITY classdefs SYSTEM "classdefs.xml">
]
>
  <collection>
&classdefs;
    <instance class="pr" id="article_0125348145.html">
      <title>press release
      </title>
      <topic>Information Retrieval</topic>
    </instance>
/.../
```

Figure A.2: A snippet of the annotation for documents.

```
<classdefs>
   <class id="ab">
     <name>Abstract</name>
       <description>
          <function>Presenting contents
          of another document</function>
        </description>
   </class>
/.../
```

Figure A.3: A snippet of the annotation for classes.

Publikationer i serien Skrifter från VALFRID

Biblioteksstudier. Folkbibliotek i flervetenskaplig belysning. Red. Romulo Enmark. (ISBN 91-971457-0-X) Skriftserien ; 1

Biblioteken och framtiden, del I. Red. Romulo Enmark. (ISBN 91-971457-1-8) Skriftserien ; 2

Biblioteken och framtiden, del II. Red. Lars Seldén. (ISBN 91-971457-2-6) Skriftserien ; 3

Hjørland, Birger: Emnerepræsentation og informationssøgning. 2. uppl. med register. (ISBN 91-971457-4-2) Skriftserien ; 4

Biblioteken, kulturen och den sociala intelligensen. Red. Lars Höglund. (ISBN 91-971457-5-0) Skriftserien ; 5

Hjørland, Birger: Faglitteratur. Kvalitet, vurdering og selektion. (ISBN 91-971457-6-9) Skriftserien ; 6

Limberg, Louise: Skolbiblioteksmodeller. Utvärdering av ett utvecklingsprojekt i Örebro län. (ISBN 91-971457-7-7) Skriftserien ; 7

Hjørland, Birger: Faglitteratur. Kvalitet, vurdering og selektion. 2. rev.udgave. (ISBN 91-971457-8-5) Skriftserien ; 8

Pettersson, Rune: Verbo-visual Communication - Presentation of Clear Messages for Information and Learning. (ISBN 91-971457-9-3) Skriftserien ; 9

Pettersson, Rune: Verbo-visual Communication - 12 Selected Papers (ISBN 91-973090-0-1) Skriftserien ; 10

Limberg, Louise: Skolbiblioteksmodeller. Utvärdering av ett utvecklingsprojekt i Örebro län. (ISBN 91-973090-1-X) Skriftserien ; 11 (nytryck av nr 7, bilaga inne i boken)

Barnbibliotek och informationsteknik. Elektroniska medier för barn och ungdomar på folkbibliotek. Red. Anette Eliasson, Staffan Lööf, Kerstin Rydsjö. (ISBN 91-973090-2-8) Skriftserien ; 12

Folkbildning och bibliotek? På spaning efter spår av folkbildning och livslångt lärande i biblioteksvärlden. Red. Maj Klasson. (ISBN 91-973090-3-6) Skriftserien ; 13

Zetterlund, Angela: Utvärdering och folkbibliotek : En studie av utvärderingens teori och praktik med exempel från folkbibliotekens förändrings- och utvecklingsprojekt (ISBN 91-973090-4-4) Skriftserien ; 14

Myrstener, Mats: På väg mot ett stadsbibliotek. Folkbiblioteksväsendets framväxt i Stockholm t o m 1927. (ISBN 91-973090-5-2) Skriftserien ; 15

Limberg, Louise: Att söka information för att lära. En studie av samspel mellan informationssökning och lärande (ISBN 91-89416-04-X, nytryck 2001 och 2003) Skriftserien ; 16

Hansson, Joacim: Om folkbibliotekens ideologiska identitet. En diskursstudie (ISBN 91-973090-7-9) Skriftserien ; 17

Gram, Magdalena: Konstbiblioteket : en krönika och en fallstudie (ISBN 91-973090-8-7) Skriftserien ; 18

Hansson, Joacim: Klassifikation, bibliotek och samhälle. En kritisk hermeneutisk studie av "Klassifikationssystem för svenska bibliotek" (ISBN 91-973090-9-5) Skriftserien ; 19

Seldén, Lars: Kapital och karriär. Informationssökning i forskningens vardagspraktik. (ISBN 91-89416-08-2, nytryck 2004) Skriftserien ; 20

Edström, Göte: Filter, raster, mönster. Litteraturguide i teori- och metodlitteratur för biblioteks- och informationsvetenskap och angränsande ämnen inom humaniora och samhällsvetenskap. (ISBN 91-89416-01-5) Skriftserien ; 21

Röster. Biblioteksbranden i Linköping. Red. Maj Klasson (ISBN 91-89416-02-3) Skriftserien ; 22

Stenberg, Catharina: Litteraturpolitik och bibliotek. En kulturpolitisk analys av bibliotekens litteraturförvärv speglad i Litteraturutredningen L 68 och Folkbiblioteksutredningen FB 80. (ISBN 91-89416-03-1) Skriftserien ; 23

Edström, Göte: Filter, raster, mönster. Litteraturguide i teori- och metodlitteratur för biblioteks- och informationsvetenskap och angränsande ämnen inom humaniora och samhällsvetenskap. Andra aktualiserade och utökade upplagan. (ISBN 91-89416-05-8) Skriftserien ; 24

Sundin, Olof: Informationsstrategier och yrkesidentiteter - en studie av sjuksköterskors relation till fackinformation vid arbetsplatsen. (ISBN 91-89416-06-6) Skriftserien ; 25

Hessler, Gunnel: Identitet och förändring - en studie av ett universitetsbibliotek och dess självproduktion. (ISBN 91-89416-07-4) Skriftserien ; 26

Zetterlund, Angela: Att utvärdera i praktiken - en retrospektiv fallstudie av tre program för lokal folkbiblioteksutveckling. (ISBN 91-89416-09-0) Skriftserien ; 27

Ahlgren, Per: The Effects on Indexing Strategy-Query Term Combination on Retrieval Effectiveness in a Swedish Full Text Database. (ISBN 91-89416-10-4) Skriftserien ; 28

Thórsteinsdóttir, Gudrun: The Information Seeking Behaviour of Distance Students. A Study of Twenty Swedish Library and Information Science Students. (ISBN 91-89416-11-2) Skriftserien ; 29

Jarneving, Bo: The Combined Application of Bibliographic Coupling and the Complete Link Cluster Method in Bibliometric Science Mapping. (ISBN 91-89416-12-0) Skriftserien ; 30

Limberg, Louise, Folkesson, Lena: Undervisning i informationssökning : Slutrapport från projektet Informationssökning, didaktik och lärande (IDOL). (ISBN 91-89416-13-9) Skriftserien ; 31

Johannisson, Jenny: Det lokala möter världen : Kulturpolitiskt förändringsarbete i 1990-talets Göteborg. (ISBN 91-89416-14-7) Skriftserien ; 32

Gärdén, Cecilia, Eliasson, Anette, Flöög, Eva-Maria, Persson, Christina, Zetterlund, Angela: Folkbibliotek och vuxnas lärande : Förutsättningar, dilemman och möjligheter i utvecklingsprojekt. (ISBN 91-89416-15-5) Skriftserien ; 33

Dahlström, Mats, Under utgivning : Den vetenskapliga utgivningens bibliografiska funktion. (ISBN 91-89416-16-3) Skriftserien ; 34

Nowé, Karen, Tensions and Contradictions in Information Management: An Activity-theoretical Approach to Information Activities in a Swedish Youth/Peace Organisations. (ISBN 978-91-85659-08-1), Skriftserien ; 35

Francke, Helena, (Re)creations of Scholarly Journals. Document and Information Architecture in Open Access Journals. (ISBN 978-91-85659-16-6), Skriftserien ; 36

Hultgren, Frances, Approaching the future: a study of Swedish school leavers' information related activities. (ISBN 978-91-89416-18-5), Skriftserien ; 37

Söderlind, Åsa, Personlig integritet som informationspolitik : Debatt och diskussion i samband med tillkomsten av Datalag (1973:289). (ISBN 978-91-89416-20-8), Skriftserien ; 38

Nalumaga, Ruth, Crossing to the Mainstream : Information Challenges an Possibilities for Female Legislators in the Ugandan Parliament. (ISBN 91-89416-20-1), Skriftserien ; 39

Johannesson, Krister, I främsta rummet : Planerandet av en högskolebiblioteksbyggnad med studenters arbete i fokus. (ISBN 978-91-89416-21-5), Skriftserien ; 40

Kawalya, Jane, The National Library of Uganda. Its inception, challenges and prospects, 1997-2007. (ISBN 91-89416-22-8), Skriftserien ; 41

Gärdén, Cecilia, Verktyg för lärande. Informationssökning och informationsanvändning i kommunal vuxenutbildning. (ISBN 978-91-98416-23-9), Skriftserien ; 42

Ponti, Marisa, Actors in Collaboration. Sociotechnical Influence on Practice-Research Collaboration. (ISBN 978-91-89416-24-6), Skriftserien ; 43

Jansson, Bertil, Bibliotekarien: om yrkets tidiga innehåll och utveckling. (ISBN 978-91-89416-25-3), Skriftserien ; 44

Olson, Nasrine, Taken for Granted. The Construction of Order in the Process of Library Management System Decision Making. (ISBN 978-91-89416-26-0), Skriftserien ; 45